On Probabilistic Methods in Fuzzy Theory

Hung T. Nguyen,[1,*] Tonghui Wang,[1,†] Berlin Wu[2,‡]
[1]*Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM 88003-8001*
[2]*Department of Mathematical Sciences, National Chengchi University, Taipei, Taiwan*

This lecture is mainly a survey of useful probabilistic methods in the theoretical analysis of fuzzy theory for modeling and design of intelligent systems. The probabilistic methods also are useful for fusing domain knowledge with numerical data in the field of intelligent data analysis. © 2004 Wiley Periodicals, Inc.§

1. INTRODUCTION

By Fuzzy theory we mean Zadeh's mathematical theory of generalized sets and their associated logics. At this state of the art, fuzzy theory has provided significant contributions to artificial intelligence via a broader umbrella of soft computing. Like other theories of mathematics, such as number theory and potential theory, fuzzy theory has its own agenda, concepts, and techniques. However, just like number theory and potential theory, probabilistic methods can be proved to be useful for fuzzy analysis. In fact, this is somewhat plausible because fuzzy theory also aims at solving decision-making problems under complex uncertainties.

In the following, because of the limited space, we will simply display the essentials of probabilistic ingredients that are useful in making fuzzy theory a quality science, i.e., providing guidelines of system designs and theoretical justifications of fuzzy inference procedures.

2. FUZZY SETS AS COVERING FUNCTIONS OF RANDOM SETS

Let U be a set. Generalizing indicator functions of (*crisp*) a subset of U, a *fuzzy subset* of U is a map $A: U \to [0, 1]$. For each $u \in U$, the value $A(u)$ is the degree of membership of u in A and hence the map A is called the membership

*Author to whom all correspondence should be addressed. e-mail: hunguyen@nmsu.edu.
†e-mail: twang@nmsu.edu.
‡e-mail: berlin@math.nccu.edu.tw.

INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS, VOL. 19, 99–109 (2004)
 Published online in Wiley InterScience (www.interscience.wiley.com). • DOI 10.1002/int.10154

function of the fuzzy subset defined by A and the ordinary (crisp) set is this. For $\alpha \in [0, 1]$, the *α-level set* of A is $A_\alpha = \{u \in U : A(u) \geq \alpha\}$. Thus,

$$A(u) = \int_0^1 A_\alpha(u) d\alpha \quad \forall\, u \in U \tag{1}$$

where we write $A_\alpha(\cdot)$ for the indicator function of the set A_α. Now, if we choose α at random, then A has the following probabilistic flavor. We view α as a random variable, defined on a probability space $(\Omega, \mathcal{A}, P)$ and uniformly distributed on [0, 1]. Then, the α-level sets are in fact *random sets,* i.e., sets obtained at random. Specifically, the map $S_A : \Omega \to \mathcal{P}(U)$ (power set of U), defined by

$$S_A(\omega) = \{u \in U : A(u) \geq \alpha(\omega)\} = A_{\alpha(\omega)}$$

is a *random element.* Now, for each fixed $u \in U$,

$$\{\omega \in \omega : u \in S_A(\omega)\} = \{\omega \in \Omega : \alpha(\omega) \leq A(u)\} \in \mathcal{A}$$

Thus, the relation in Equation 2.1 is replaced by the following relation between a fuzzy set A and the random set S_A:

$$A(u) = P(u \in S_A) \quad \forall\, u \in U \tag{2}$$

In more formal terms, if we specify a σ-field $\mathcal{U}$ of subset of $\mathcal{P}(U)$, then a random set S is a map S: $\Omega \to \mathcal{P}(U)$, which is $\mathcal{A}$-$\mathcal{U}$ measurable (a *random element* is a map from Ω to an abstract space E, which is $\mathcal{A}$-$\mathcal{E}$ measurable, where $\mathcal{E}$ is a σ-field on E). In the case in which $\{\omega \in \Omega$: $u \in S(\omega)\} \in \mathcal{A}$ for any $u \in U$, as in the case of S_A shown previously, the map

$$u \in U \;\to\; \pi_S(u) = P(u \in S)$$

is called the *covering function* of the random set S. Obviously, a covering function of a random set defines a membership function, i.e., a fuzzy subset of U. For a history of this connection between fuzzy sets and random sets, see, e.g. Ref. 1.

Several interesting observations can be drawn from the relation in Equation 2.

2.1. Fuzziness Is a Weakened Form of Randomness

Without going into philosophical issues, we stay with the mathematics! The randomness of a random set S is modeled by its distribution, i.e., by the probability measure (law) $P_S = PS^{-1}$ on $\mathcal{U}$; from it $\pi_S(\cdot)$ can be determined. Now, if we are able only to specify $\pi_S(\cdot)$, can we deduce P_S? Except in special cases (see Section 2.2), the answer is no. This is somewhat similar to the *moment problem* of random variables. To see this, we recall the following formula.[2]

Let S be a random set, defined on $(\Omega, \mathcal{A}, P)$, with values in $\mathcal{C} \subseteq \mathcal{B}(\Re^d)$ (Borel subsets of $\Re^d$). Let $\sigma(\mathcal{C})$ be a σ-field on $\mathcal{C}$. Let μ denote the Lebesgue measure on $\Re^d$. Suppose that $\mu(S)$ is a random variable. Then, the moments of $\mu(S)$ can be obtained from the knowledge of *multiple* covering functions of S. Specifically, for each $n \geq 1$, let $\pi_n : \Re^{nd} \to [0, 1]$ be defined by

$$\pi_n(u_1, \ldots, u_n) = P(\omega : \{u_1, \ldots, u_n\} \in S(\omega))$$

where $u_i \in \mathfrak{R}^d$, $i = 1, \ldots, n$. Then, under some measureability conditions,

$$E[\mu(S)]^n = \int_{\mathfrak{R}^{nd}} \pi(u_1, \ldots, u_n)(\otimes_{i=1}^{n})(d\mu_1, \ldots, d\mu_n)$$

where $\otimes_{i=1}^{n}$ is the product measure $\mu \otimes \mu \otimes \cdots \otimes \mu$ on $\mathfrak{R}^{nd}$.

2.2. Choquet Capacities

Unlike random elements taking values in Banach spaces, it seems that the analysis of random sets is somewhat delicate, perhaps because of its set nature. The complete theory of random sets is formulated for *random closed sets* on $\mathfrak{R}^d$ or, more generally, on local compact, Hausdorff, separable topological spaces (see, e.g. Ref. 3). For such random sets, their probability laws are determined completely by their *capacity functionals* (Choquet capacities) via the *Choquet theorem,* namely, Let $\mathcal{K}$ and $\mathcal{F}$ denote the classes of compact and closed sets of $\mathfrak{R}^d$, respectively. Let $\mathcal{B}(\mathcal{F})$ denote the Borel σ-field on $\mathcal{F}$ generated by the so-called "hit-and-miss" topology of $\mathcal{F}$. Then, a set function $T : \mathcal{K} \to [0, 1]$ determines uniquely a probability P_S on $\mathcal{B}(\mathcal{F})$ such that

$$T(K) = P_S\{F \in \mathcal{F} : F \cap K \neq \varnothing\} \quad \forall K \in \mathcal{K}$$

if and only if T satisfies

(a) $T(\varnothing) = 0$

(b) If $K_n \downarrow K$ in $\mathcal{K}$, then $T(K_n) \downarrow T(K)$

(c) T is alternating of infinite order, i.e., T is monotone increasing [i.e., $K_1 \subseteq K_2$ implies that $T(K_1) \leq T(K_2)$] and for any $n \geq 2$, $K_1, \ldots K_n \in \mathcal{K}$,

$$T\left(\bigcap_{i=1}^{n} K_i\right) \leq \sum_{\varnothing \neq I \subseteq \{1, \ldots, n\}} (-1)^{|I|+1} T\left(\bigcup_{i \in I} K_i\right)$$

where $|I|$ denotes the cardinality of the set I.

Here is a situation in which a fuzzy subset A of $\mathfrak{R}^d$ determines the distribution (i.e., the capacity functional) of the random set S_A. For S_A to be a random *closed* set, it suffices to suppose that $A(\cdot)$ is upper semicontinuous (USC, i.e., $\forall t \in \mathfrak{R}$, the set $\{x \in \mathfrak{R}^d : A(x) \geq t\}$ is a closed set of $\mathfrak{R}^d$.

Recall that $S_A(\omega) = A_{\alpha(\omega)}$. Let $K \in \mathcal{K}$, and then because A is USC, we have

$$\begin{aligned} P(S_A(\omega) \cap K \neq \varnothing) &= P(\omega : \alpha(\omega) \leq A(x) \text{ for some } x \in K) \\ &= P(\omega : \alpha(\omega) \leq \sup_{x \in K} A(x)) = \sup_{x \in K} A(x) = T(K) \end{aligned}$$

Thus, it suffices to show that T satisfies clauses (a)–(c) in Choquet theorem. Clause a is obvious and Clause b is left as an exercise! What is interesting is that Clause

c follows from a general property of set functions called *maxitivity* (a property shared by *possibility measures*). Indeed, observe that the foregoing T is *maxitive,* i.e., $T(K_1 \cup K_2) = \max(T(K_1), T(K_2))$. And any maxitive set functions are alternating of infinite order (for a proof, see e.g., Ref. 4).

2.3. A Connection with Fuzzy Logics

The concept of *t-norms* is used in fuzzy logics to model the logical connective "AND." They are related to the concept of *copulas* in probability and statistics (see, e.g., Ref. 5). Roughly speaking, copulas are functions that combine marginal distributions into joint distributions of random vectors (see, e.g., Ref. 6). Specially, if F is a distribution function on $\Re^d$ with marginals $F_1, \ldots, F_d$, then there exists an n-copula $C : [0, 1]^d \to [0, 1]$ such that

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \quad \forall (x_1, \ldots, x_d) \in \Re^d$$

For example, copulas are t-norms if and only if they are associative. Copulas can be used to show that the random set S_A associated with the fuzzy set A is *canonical* in the sense that S_A is a *nested* random set. Specifically, among all possible random sets in which their covering functions are precisely A, S_A is the only nested random set representing A. For a proof, see, e.g., Ref. 7.

3. FUZZY SETS AS POSSIBILITY DISTRIBUTIONS

Among many situations in which membership functions of fuzzy concepts (in a natural language) can be interpreted as distributions of variables, perhaps *fuzzy control* provides the most plausible one. Fuzzy control is an example of successful industrial applications in which almost all concepts developed in fuzzy theory (except perhaps fuzzy measures) are in use, especially inference procedures. Fuller[8] gives an excellent introduction to neurofuzzy control.

Consider a controlled system where the input variable $\mathbf{x} = (x_1, \ldots, x_d)$ is taking values in $\mathcal{X} \subseteq \Re^d$, $\mathcal{X} = \otimes_{i=1}^d \mathcal{X}_i$, and the output variable y is taking values in $\mathcal{Y} \subseteq \Re$.

By a *fuzzy rule base* $\mathcal{R}$, we mean a collection of "If $\cdots$, Then $\cdots$," rules in natural language modeled by fuzzy sets of the form

$$R_j: \quad \text{If } x_1 \text{ is } A_{j1}, \ldots, x_d \text{ is } A_{jd} \text{ then } y \text{ is } B_j, j = 1, \ldots, k$$

where A_{ji}'s and B_j's are fuzzy subsets of $\mathcal{X}_i$ and $\mathcal{Y}$, respectively.

In standard fuzzy control, the purpose is to infer a control law $y = f(\mathbf{x})$ from $\mathcal{R}$. Now, setting $A_j = \Pi_{i=1}^d A_{ji}$, i.e.,

$$A_j(x_1, \ldots, x_d) = t(A_{j1}(x_1), \ldots, A_{jd}(x_d))$$

where t is some t-norm, we see that $\mathcal{R}$ is a data set of the form (A_j, B_j), $j = 1, \ldots, k$ of pairs of fuzzy numbers. Each pair (A_j, B_j), generalizing numerical data, is a *relation* between $\mathbf{x}$ and y, rather than a *causality.* In control context, B_j is a recommendation for action where the input is A_j rather than A_j is the cause of

B_j. In other words, one should view each $\mathcal{R}_j$ as a *fuzzy relation* on $\mathcal{X} \times \mathcal{Y}$ of a general nature rather than interpreting literally "if $\cdots$, then $\cdots$," as an (fuzzy) implication operator.

Remark. The situation might be different in some domain of expert systems. For example, in medical expert systems, a rule $\mathcal{R}$ might express a causal relationship and hence could be modeled by an implication operator.

Let R_j be the relation on $\mathcal{X} \times \mathcal{Y}$ determined by

$$R_j(\mathbf{x}, y) = s(A_j(\mathbf{x}, B_j(y))$$

where s is some function from $[0, 1]^2 \to [0, 1]$. Thus, for an observed input $\mathbf{x}$, the rule $\mathcal{R}_j$ produces a fuzzy subset of $\mathcal{Y}$: $y \to R_j(\mathbf{x}, y)$, $y \in \mathcal{Y}$. The meaning of this membership function is clear: it is information on possible values of y to be taken, in other words, it is a *possibility distribution* on the output y. If we want to obtain a specific value $y(\mathbf{x})$, then we can *defuzzify* $R_j(\mathbf{x}, \cdot\ )$, i.e., summarizing the possibility distribution $R_j(\mathbf{x}, \cdot\ )$ into a single value. This can be done in several ways. Now, a possibility distribution needs not be a *probability density function,* but it can be transformed into such a function, say, by normalizing. This reminds us of various *probabilistic proofs* in analysis, e.g., that of the Weierstrass theorem on approximating continuous functions on intervals by polynomials. The popular defuzzification method in fuzzy control, namely, the center-of-gravity method, is a form of *statistical expectation* or, more specifically, a *conditional mean* given the input $\mathbf{x}$. Note also that in Sugeno's model, the overall output is a conditional mean of a *fuzzy random variable.*

It is interesting to note at this point that the concept of *possibility measures,* which are derived from the concept of possibility distributions also has some probability interpretation in terms of random sets. For example, let S be a random set taking values as subsets of a finite set U. Suppose that the values of S form a nested collection of subsets of U, i.e., $A_1 \subseteq \cdots \subseteq A_n \subseteq U$ with $\alpha_i = P(S = A_i)$, $\sum_{i=1}^{n} \alpha_i = 1$. Then, ϕ: $\mathcal{P}(U) \to [0, 1]$ defined by $\phi(A) = 1 - \sum_{A_i \subseteq A'} \alpha_i$ (where A' denotes the set complement of A in U) is a possibility measure, i.e., $\phi(\varnothing) = 0$, $\phi(U) = 1$, and for any family $\{B_j : j \in J\}$ of subsets of U we have

$$\phi(\cup_{j \in J} B_j) = \max\{\phi(B_j) : j \in J\}$$

where ϕ is a maxitive set function.

This is in fact a *well-known general result*,[9] namely, an *upper probability* is a possibility measure if and only if its focal elements are nested.

Remark. In reasoning with imprecise probabilities, the framework is this. The true law P_0 on $(\Omega, \mathcal{A})$ is only known to belong to a class of probability measures $\mathcal{P}$. The upper and lower probabilities are

$$G(A) = \sup\{P(A) : P \in \mathcal{P}\}, \quad F(A) = \inf\{P(A) : P \in \mathcal{P}\}$$

When Ω is finite and $\mathcal{P} = \{P : F \leq P\}$, then F is the distribution function of some random set S in which its probability "density" f is the Mobius transform of F: $f(A) =$

$\sum_{B \subseteq A}(-1)^{|A-B|}F(B)$. Such F's are called *belief functions* and the dual $G(A) = 1 - F(A')$ is called a plausibility function. For a discussion on the case where Ω is infinite, see Ref. 10. The focal elements of F are nothing more than the support of S.

4. VAPNIK-CHERVONENKIS DIMENSION

When using fuzzy theory in intelligent technologies, it is necessary to assess the capability and the performance of fuzzy systems (control or expert systems). After all, these are *learning machines.* As such, the general framework of *statistical learning theory* is useful. This is exemplified by *neural networks* (see, e.g., Ref. 11), in which concepts from statistics are used in evaluating generalization capacity, in particular, the concept of *Vapnik-Chervonenkis dimension* (VC-dimension) is essential (see, e.g., Ref. 12). To the best of our knowledge, VC-dimension of classes of functions still has not been used in the study of fuzzy systems. In forthcoming work, we will detail our findings. Here, we simply indicate the setup.

In a simple design of a fuzzy controlled system, the inference mechanism can be put in a statistical learning setting. Indeed, with *fuzzy partitions* of input and output spaces modeled by triangular membership functions and with appropriate chosen t-norms and t-conorms for logical connectives involved in the fuzzy rules, the derivation of a successful control law is nothing more than choosing a function f_0 in a (parametric) class $\mathcal{F}$ of functions. To judge the performance of f_0, we need to specify a criterion, just like in the case of neural networks where the *back-propagation algorithm* is judged with respect to the minimization of a risk functional. In this context, important issues such as control capacity and performance need to be addressed. Mathematically, these are related to the problem of convergence of the algorithm and its rate of convergence.

Basically, we need to inquire about the "size" of the class $\mathcal{F}$ of functions in a fuzzy system design. The following concept of dimension or index due to Vapnik and Chervonenkis in pattern recognition is essential. Here, we choose to introduce this concept to researchers in fuzzy theory.

Stochastic processes are random elements that take values in functional spaces such as $C[0, 1]$, the Banach space of continuous functions defined on the unit interval (sample paths of Brownian motion). Thus, classical the framework of Euclidean space $\mathfrak{R}^d$ should be extended to arbitrary spaces. Roughly speaking, a *random element* X is defined as follows.

Let $(\Omega, \mathcal{A})$ and $(\mathcal{X}, \mathcal{B})$ be two measurable spaces. A random element X is a map $\Omega \to \mathcal{X}$, which is $\mathcal{A}$-$\mathcal{B}$ measurable. If P is a probability measure on $(\Omega, \mathcal{A})$, then the *law* of X is the induced probability measure $X = PX^{-1}$ on $(\mathcal{X}, \mathcal{B})$. The fundamental problem in statistics is this. Can we estimate the unknown law P_X from the random sample $X_1, \ldots, X_n$ drawn from X?

This problem was solved in the case of $\mathfrak{R}$ or $\mathfrak{R}^d$. Essentially, this is because of the structure of $\mathfrak{R}$. First, in view of the Lebesgue-Stieljes measure theory, each law P_X on $(\mathfrak{R}, \mathcal{B}(\mathfrak{R}))$ can be identified with a *distribution function* F: $\mathfrak{R} \to [0, 1]$, $F(x) = P(X \leq x)$, Thus, we consider only the problem of estimating a function rather than a measure. Second, because $\mathfrak{R}$ is separable, an uncountable supremum of random variables is a random variable.

The solution is the celebrated *Glivenko-Cantelli theorem* (1933); let the *empirical distribution function* $F_n(x) = 1/n \sum_{j=1}^{n} I_{(-\infty,x]}(X_j)$, where I_A denotes the *indicator function of the set (event)* A. Then, $F_n(x) \to F(x)$, as $n \to \infty$, almost surely (a.s.) and uniformly in $x \in \Re$, i.e., $P(\sup_x |F_n(x) - F(x)| \to 0) = 1$.

To consider the general setup where $(\mathcal{X}, \mathcal{B})$ is arbitrary, we need to use the measure (laws) at the place of distribution functions. Now, observe that

$$F_n(x) = \mathbf{F}_n(f) = \int_{\Re} f(y) d\mathbf{F}_n(y)$$

where $\mathbf{F}_n$ denotes the law associated with the distribution function F_n and $f(y) = I_{(-\infty,x]}(y)$. We also can write $\mathbf{F}_n = 1/n \sum_{j=1}^{n} \delta_{X_j}$, where δ_{X_j} is the (random) Dirac measure $\delta_{X_j}(A) = I_A(X_j)$, $A \in \mathcal{B}$. Similarly, $F(x) = \mathbf{F}(f)$. Thus, instead of $f(y) = I_{(-\infty,x]}(y)$, we can take any arbitrary measurable real-valued function f defined on $\mathcal{X}$ and consider the population $f(X)$ to investigate uniform laws of large numbers in the general setting provided that $Ef(X)$ is finite. Note also that the uniform version of Glivenko-Cantelli theorem is with respect to the *class* $\mathcal{C}$ *of subsets* $(-\infty, x]$, $x \in \Re$, or by identification, the *class* $\mathcal{F}$ *of indicator functions* $I_{(-\infty,x]}$, $x \in \Re$. Thus, we have, as $n \to \infty$,

$$\|\mathbf{F}_n - \mathbf{F}\|_{\mathcal{C}} \to 0, \quad \text{a.s.} \quad \text{or} \quad \|\mathbf{F}_n - \mathbf{F}\|_{\mathcal{F}} \to 0, \quad \text{a.s.}$$

where $\|\mathbf{F}\|_{\mathcal{C}} = \sup_{C \in \mathcal{C}} |\mathbf{F}(C)|$ and $\|\mathbf{F}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbf{F}(f)|$. Of course, for general $\mathcal{X}$, we need to worry about the measureability of the quantity $\|\mathbf{F}_n - \mathbf{F}\|_{\mathcal{F}}$.

To have Glivenko-Cantelli result in a general setting, we need to find *sufficient conditions* for it.

Let $\mathcal{C}$ be a class of sets of $\mathcal{X}$. For any *finite* subset A of $\mathcal{X}$, let $\#c(A) = \#\{A \cap C : C \in \mathcal{C}\}$, i.e., the cardinality of the set of subsets of A picked out by $\mathcal{C}$. If $\#c(A) = 2^{\#(A)}$, then we say that $\mathcal{C}$ shatters A, i.e., $\mathcal{C}$ picked out *all* subsets of the finite set A.

The *growth function* of $\mathcal{C}$ is defined to be

$$g_{\mathcal{C}}(n) = \max\{\#c(A) : \#(A) = n\}, \quad n = 0, 1, 2, \ldots$$

Note that $g_{\mathcal{C}}(n) \leq 2^n$. The VC-*dimension* of $\mathcal{C}$ is defined as

$$D(\mathcal{C}) = \begin{cases} \infty & \text{if } \{n : g_{\mathcal{C}}(n) < 2^n\} = \varnothing \\ \inf\{n : g_{\mathcal{C}}(n) < 2^n\} & \text{if not} \end{cases}$$

Thus, $D(\mathcal{C})$ is the smallest integer n for which no set of cardinality n is shattered by $\mathcal{C}$.

Examples.

(i) $\mathcal{C} = \{(-\infty, x] : x \in \Re\}$ has VC-dimension two, because $\mathcal{C}$ can shatter sets of one element but can not shatter *any* one set of two elements. This situation is general: any class $\mathcal{C}$ of subsets (with at least two elements) of an arbitrary set $\mathcal{X}$ that is *nested* has $D(\mathcal{C}) = 2$.

(ii) $\mathcal{C} = \{(y, x] : x, y \in \Re\}$ has VC-dimension three because $\mathcal{C}$ can not pick out the subset $\{x, z\}$ of any set $\{x, y, z\}$, where $x < y < z$.
(iii) $\mathcal{X} = \Re^2$, $\mathcal{C} = \{\text{all closed disks}\}$ has VC-dimension four.
(iv) $\mathcal{X} = \Re^d$, $\mathcal{C} = \{\text{all convex subsets}\}$ has VC-dimension ∞.

To extend the concept of VC-dimension for classes of sets or, equivalently, for classes of indicator functions to classes of real-valued functions, we can use various ways to identify functions with collections of a set.

For example, we can use the concept of a *subgraph* of a function f, namely, the subset $S(f) = \{(x, t) : t < f(x)\} \subset \mathcal{X} \times \Re$, and define the VC-dimension of a class of function $|calF$ to be the VC-dimension of its class of subgraphs $\{S(f) : f \in \mathcal{F}\}$.

The general statistical problem is this. Given a random sample $X_1, \ldots, X_n$ drawn from X with unknown *law* Q on $(\mathcal{X}, \mathcal{B})$, we wish to estimate the law Q, e.g., by the sequence of empirical measures Q_n, which is defined by $Q_n(A) = 1/n \sum_{j=1}^{n} I_A(X_j)$. Then, for the estimation to be useful, we need at least two things:
(i) Q_n is *uniformly consistent* for Q, i.e., $\sup_{A \in \mathcal{B}} |Q_n(A) - Q(A)| \to 0$ a.s.
(ii) The rate of the uniform convergence in Item (i)

However, in general, even Item i is not satisfied for arbitrary Q and $\mathcal{B}$. For small $\mathcal{B}$ (finite $\mathcal{B}$), of course, Item (i) holds; because

$$\sup_{A \in \mathcal{B}} |Q_n(A) - Q(A)| \leq \sum_{A \in \mathcal{B}} |Q_n(A) - Q(A)| \to 0, \quad \text{a.s.}$$

by the strong law of large numbers, the sum being finite.

In the other extreme (see, e.g., Ref. 12), consider $(\mathcal{X}, \mathcal{B}) = ((0, 1), \mathcal{B}_1)$ with $Q = dx$ [Lebesgue measure on (0, 1)]. Let $X_1, \ldots, X_n$ be drawn according to dx. For any $\varepsilon > 0$, we can find an event $A^* \in \mathcal{B}_1$ such that $Q_n(A^*) = 1$ and $dx(A^*) < \varepsilon$ (take A^* to be the union of n small intervals A_j each centered at X_j, of length $< \varepsilon/n$).

Consequently, for any n, $P(\sup_{A \in \mathcal{B}_1} |dx(A) - Q_n(A)| = 1) = 1$. Thus, in this case, Item (i) does not hold.

Therefore, the general problem of statistics should be formulated as follows:
(a) For which Q is the Foregoing Item (i) possible?
(b) If Item (i) is not possible, can we obtain a "partial" result, i.e., determine subclasses $\mathcal{C}$ (not necessarily sub–σ-algebras) of $\mathcal{B}$ such that Item (i) holds on $\mathcal{C}$? In this case, we say that Q_n provides partial uniform convergence to Q determined by $\mathcal{C}$. We have seen that partial uniform convergence can take place when uniform convergence fails (Glivenko-Cantelli theorem).

For Question (b), it turns out that the condition $D(\mathcal{C}) < \infty$ is *sufficient* (VC theorem). The condition clearly is not necessary as the foregoing example shows: $D(\mathcal{B}_1) = \infty$. Note that if we consider a subclass $\mathcal{C}$ of $\mathcal{B}_1$, e.g., $\mathcal{C} = \{(0, x] :, x \in (0, 1)\}$, which is nested, and then Question (b) holds with $D(\mathcal{C}) = 2$. Moreover, we need the value $D(\mathcal{C})$ to specify the rate of convergence.

When $D(\mathcal{C}) = n < \infty$, then $n - 1$ is the largest cardinality of a set shattered by $\mathcal{C}$, specifically, $D(\mathcal{C}) - 1 = \sup\{n : g_{\mathcal{C}}(n) = 2^n\}$.

Noting that $g_{\mathcal{C}}(n) = 2^n$ means that there exists a set of size n that can be shattered by $\mathcal{C}$. Note also that $D(\mathcal{C}) = 0$ if and only if $\mathcal{C}$ is empty and $\mathcal{C} \subset \mathcal{D}$ implies $D(\mathcal{C}) \leq D(\mathcal{D})$.

Now, let us look at $D(\mathcal{F})$ for a class of functions $\mathcal{F}$. Recall that by definition, $D(\mathcal{F}) = D(S(\mathcal{F}))$.

Examples.

(i) Let $\phi : \Re \to \Re$ be monotone, e.g., nonincreasing, and let $\mathcal{F}$ be the class of functions $f_h = \phi(x - h)$, for $h \in \Re$. Then, $D(\mathcal{F}) = 2$. Indeed, for any $h \leq h'$, $\phi(x - h) \leq \phi(x - h')$ so that the subgraphs are nested: $S(f_h) \subseteq S(f_{h'})$, i.e., $S(\mathcal{F})$ is a class of nested subsets of $\Re^2$ and, hence, has $D(\mathcal{F}) = D(S(\mathcal{F})) = 2$.

(ii) Let $\mathcal{F}$ be a finite dimensional real vector space of real-valued and measurable functions defined on $\mathcal{X}$, e.g., $\dim(\mathcal{F}) = d < \infty$. Let $\{(x_i, t_i), i = 1, \ldots, n\} = A$ be a subset in $\mathcal{X} \times \Re$ of cardinality n. Consider the class $\mathcal{P}$ of vectors in $\Re^n$ of the form $\{(f(x_1) - t_1, \ldots, f(x_n) - t_n) : f \in \mathcal{F}\}$. Then, there exists a nonzero vector $\mathbf{a} = (a_1, \ldots, a_n) \in \Re^n$ orthogonal to $\mathcal{P}$. Of course, this vector $\mathbf{a}$ will depend on the x_i's and t_i's but not on f. It suffices to take $\mathbf{a}$ to be a solution of the system of $d + 1 = n - 1$ linear equations for n variables a_i, $i = 1, \ldots, n$:

$$\sum_{i=1}^{n} a_i t_i = 0, \quad \sum_{i=1}^{n} a_i g_j(x_i) = 0, \quad j = 1, \ldots, d$$

where g_j, $j = 1, \ldots, d$ is a basis for $\mathcal{F}$ so that any $f \in \mathcal{F}$ is of the form $f = \sum_{j=1}^{d} c_j g_j$. Thus, taking such an $\mathbf{a}$ with at least one component $a_i > 0$, we have, for any $f \in \mathcal{F}$, $\sum_{i=1}^{n} a_i(f(x_i) - t_i) = 0$ or, equivalently,

$$\sum_{a_i > 0} a_i(f(x_i) - t_i) = -\sum_{a_i < 0} a_i(f(x_i) - t_i) \tag{3}$$

Then, the subset $\{(x_i, t_i) : \text{with } i \text{ such that } a_i > 0\}$ of A can not be picked out by $S(\mathcal{F})$, i.e., this subset is not of the form $\{(x_i, t_i) : \text{such that } t_i < f(x_i)\}$ for some f in $\mathcal{F}$. Because if it is so, then the left side of Equation 3 is strictly positive, the right side is nonpositive (with the convention that the sum over an empty set is zero). Thus, $D(\mathcal{F}) \leq d + 2$.

5. FUSION OF FUZZY DOMAIN KNOWLEDGE WITH STATISTICAL DATA

The use of probabilistic methods is apparent in building intelligent systems when we wish to take into account all available information, *in any form,* to reach better decisions. This is the essence of the new field of *intelligent data analysis*

(see, e.g., Ref. 13). One form of this analysis is the problem of combining domain knowledge with measurement data. In fuzzy control or expert systems, this will happen when, e.g., we have a linguistic (fuzzy) rule base *and* numerical measurements (x_i, y_i), $i = 1, \ldots, m$ on the behavior of the system. To combine these two different types of data (linguistic and numerical), we need to transform one type to the other. Because linguistic information is richer, it is desirable to transform numerical data into linguistic rules. This is exemplified by procedures known in fuzzy control as extracting fuzzy rules from learning examples.

We will give an example for fusing domain knowledge with statistical data in the framework of *Bayesian statistics.*

As testified by the recent workshop on combining data with domain knowledge at Stanford University, June 2000, this problem is of central importance in intelligent decision making where one should use all kinds of information and numerical information from measurements as well as perception, to reach better decisions in complex situations. Domain knowledge is expressed in general linguistic terms, and as such, fuzzy modeling is necessary. In fact, this is precisely why fuzzy control is considered in the first place; one common and practical aspect to domain knowledge is the behavior of a system under consideration. This behavior is captured by a set of fuzzy rules on training samples. Here, to illustrate the probabilistic method, we look at a simple example.

Consider the case in which the numerical information comes from the observed values of a random variable X, which follows, e.g., a parametric model $\{f(x, \theta) : \theta \in \Theta\}$. Suppose we wish to estimate the true parameter θ_0 with the following data.

(i) A random sample $(X_1, \ldots, X_n)$ drawn from X
(ii) Domain knowledge: some additional information about θ_0 before performing the experiment

This situation is idealistic for Bayesian statistics if the additional information is modeled as a prior distribution of Θ viewed as a random variable. Now, consider the case in which the additional information about θ_0 is linguistic of the form "θ_0 is small," where we model the linguistic label "small" as a fuzzy subset with membership function A on, e.g., $\Re^+$. Note that if several experts give different opinions about θ_0, then we will use the fuzzy logic connective "and" (via some t-norm operator) to combine them into a single fuzzy set.

The question of interest is how to fuse the nonstatistical data, namely, the membership function A (not a probability density function) with the statistical data in order to better estimate θ_0?

One obvious way is to transform this situation into a Bayesian framework by simply normalizing A to obtain a bona fide prior probability density function: $\phi(\theta) = A(\theta)/\int_{\Re^+} A(\lambda)d\lambda$.

Now, we show that this normalization procedure does have a probabilistic interpretation in terms of random sets. To be simple, suppose the parameter space Θ is finite with cardinality N. Let $Y : \Omega \to \Theta$ be uniformly distributed. We show that $\phi(\theta)$ is nothing more than the updated version of a uniform prior on Θ by experts' opinion A. Indeed, let S denote the random set associated with A. Then,

$$\frac{A(\theta)}{\sum_{\lambda\in\Theta} A(\lambda)} = \frac{A(\theta)/N}{\sum_{\lambda\in\Theta A(\lambda)/N}} = \frac{P(\omega : Y(\omega) = \theta \in S(\omega))}{P(\omega : Y(\omega) \in S(\omega))} = P(Y = \theta | Y \in S)$$

assuming Y and S are independent.

Here is another example where fuzzy information can be combined with statistical data in a probabilistic setting. Consider the normal density function

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$$

The likelihood that we observed the statistical data x (e.g., for sample of size $n = 1$) and experts assigned $A(\theta)$ as the degree to which θ is compatible with A is

$$L(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) A(\theta)$$

The maximum likelihood estimator of θ_0 is obtained by maximizing $L(x, \theta)$ or, equivalently, minimizing $-\log L(x, \theta)$ over $\theta \in \Theta$, which is $\hat{\theta} = x + \sigma^2 A'(\theta)/A(\theta)$, exhibiting the correcting term $\sigma^2 A'(\theta)/A(\theta)$ because of the additional information A given by experts. Note that a random size n in the foregoing estimator is replaced by the sample mean.

References

1. Goodman IR, Nguyen HT. Uncertainty models for knowledge-based systems. New York: North Holland; 1985.
2. Robbins HE. On the measure of random set. Ann Math Stat 1944;15:70–74.
3. Matheron G. Random sets and integral geometry. New York: Wiley; 1975.
4. Goodman IR, Mahler R, Nguyen HT. Mathematics of data fusion. New York: Kluwer Academic Press; 1997.
5. Nguyen HT, Walker EA. A first course in fuzzy logic, 2nd edition. Chapman & Hall/CRC; Boca Raton, FL, 2000.
6. Nelsen RB. An introduction to copulas. Lecture Notes in Statistics, No. 39, New York: Springer; 1999.
7. Nguyen HT. Some mathematical structures for computational information. J Inf Sci 2000;128:67–89.
8. Fuller R. Introduction to neuro-fuzzy systems. New York: Springer; 2000.
9. Dubois D, Prade H. When upper probabilities are possibility measures? J Fuzzy Set Syst 1992;49:65–74.
10. Nguyen HT. Fuzzy measures and related topics. In: Proc of the Workshop on Current Trends and Developments in Fuzzy Logic, Thessaloniki, Greece, October 16–20, 1998. pp 63–99.
11. Fine TL. Feed forward neural network methodology. New York: Springer; 1999.
12. Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
13. Berthold M, Hand DJ. Intelligent data analysis. New York: Springer; 1999.