To appear in the *IEEE Transactions on Systems, Man, and Cybernetics: Part B*
Current Draft: 20 January 2011

Bayesian Inference with Adaptive Fuzzy Priors and Likelihoods

Osonde Osoba, Sanya Mitaim, and Bart Kosko

***Abstract*—Fuzzy rule-based systems can approximate prior and likelihood probabilities in Bayesian inference and thereby approximate posterior probabilities. This fuzzy approximation technique allows users to apply a much wider and more flexible range of prior and likelihood probability density functions than found in most Bayesian inference schemes. The technique does not restrict the user to the few known closed-form conjugacy relations between the prior and likelihood. It allows the user in many cases to describe the densities with words. And just two rules can absorb any bounded closed-form probability density directly into the rulebase. Learning algorithms can tune the expert rules as well as grow them from sample data. The learning laws and fuzzy approximators have a tractable form because of the convex-sum structure of additive fuzzy systems. This convex-sum structure carries over to the fuzzy posterior approximator. We prove a uniform approximation theorem for Bayesian posteriors: An additive fuzzy posterior uniformly approximates the posterior probability density if the prior or likelihood densities are continuous and bounded and if separate additive fuzzy systems approximate the prior and likelihood densities. Simulations demonstrate this fuzzy approximation of priors and posteriors for the three most common conjugate priors (as when a beta prior combines with a binomial likelihood to give a beta posterior). Adaptive fuzzy systems can also approximate non-conjugate priors and likelihoods as well as approximate hyperpriors in hierarchical Bayesian inference. The number of fuzzy rules can grow exponentially in iterative Bayesian inference if the previous posterior approximator becomes the new prior approximator.**

I. Bayesian Inference with Fuzzy Systems

Additive fuzzy systems can extend Bayesian inference because they allow users to express prior or likelihood knowledge in the form of if-then rules. Fuzzy systems can approximate *any* prior or likelihood probability density functions (pdfs) and thereby approximate *any* posterior pdfs. This allows a user to describe priors with fuzzy if-then rules rather than with closed-form pdfs. The user can also train the fuzzy system with collateral data to adaptively grow or tune the fuzzy rules and thus to approximate the prior or likelihood pdf. A simple two-rule system can also exactly represent a bounded prior pdf if such a closed-form pdf is available. So fuzzy rules *substantially* extend the range of knowledge and statistical structure that prior or likelihood pdfs can capture–and they do so in an expressive linguistic framework based on multivalued or fuzzy sets [33].

Figure 1 shows how five tuned fuzzy rules approximate the skewed beta prior pdf $\beta(8,5)$. Learning has sculpted the five if-part and then-part fuzzy sets so that the approximation is almost exact. Users will not in general have access to such training data because they do not know the functional form of the prior pdf. They can instead use any noisy sample data at hand or just state simple rules of thumb in terms of fuzzy sets and thus implicitly define a fuzzy system approximator F. The following prior rules define such an implied skewed prior that maps fuzzy-set descriptions of the parameter random variable Θ to fuzzy descriptions $F(\Theta)$ of the occurrence probability:

Rule 1: If Θ is *much smaller* than $\frac{1}{2}$ then $F(\Theta)$ is *very small*
Rule 2: If Θ is *smaller* than $\frac{1}{2}$ then $F(\Theta)$ is *small*
Rule 3: If Θ is *approximately* $\frac{1}{2}$ then $F(\Theta)$ is *large*
Rule 4: If Θ is *larger* than $\frac{1}{2}$ then $F(\Theta)$ is *medium*
Rule 5: If Θ is *much larger* than $\frac{1}{2}$ then $F(\Theta)$ is *small*

Learning shifts and scales the Cauchy bell curves that define the if-part fuzzy sets in Figure 1. The tuned bell curve in the third rule has shifted far to the right of the equi-probable value $\frac{1}{2}$. Different prior rules and fuzzy sets will define different priors just as will different sets of sample data. The simulations results in Figures 3-11 show that such fuzzy rules can quickly learn an implicit prior if the fuzzy system has access to data that reflects the prior. These simulations give probative evidence that an informed expert can use fuzzy sets to express reasonably accurate priors in Bayesian inference even when no training data is available. The uniform fuzzy approximation theorem in [13], [15] gives a theoretical basis for such rule-based approximations of priors or likelihoods. Theorem 2 below further shows that such uniform fuzzy approximation of priors or likelihoods leads in general to the uniform fuzzy approximation of the corresponding Bayesian posterior.

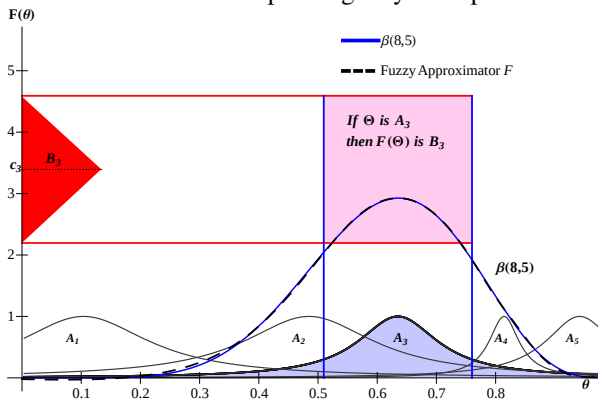

Fig. 1. Five fuzzy if-then rules approximate the beta prior $h(\theta) = \beta(8,5)$. The five if-part fuzzy sets are truncated Cauchy bell curves. An adaptive Cauchy SAM (Standard Addtive Model) fuzzy system tuned the sets' location and dispersion parameters to give a nearly exact approximation of the beta prior. Each fuzzy rule defines a patch or 3-D surface above the input-output planar state space. The third rule has the form "If $\Theta = A_3$ then B_3" where then-part set B_3 is a fuzzy number centered at centroid c_3. This rule might have the linguistic form "If Θ is *approximately* $\frac{1}{2}$ then $F(\Theta)$ is *large*." The training data came from 500 uniform samples of $\beta(8,5)$. The adaptive fuzzy system cycled through each training sample 6,000 times. The fuzzy approximator converged in fewer than 200 iterations. The adaptive system also tuned the centroids and areas of all five then-part sets (not pictured).

Bayesian inference itself has a key strength and a key weakness. The key strength is that it computes the posterior pdf $f(\theta|x)$ of a parameter θ given the observed data x. The posterior pdf gives all probabilistic information about the parameter given the available evidence. The key weakness is that this process requires that the user produce a prior pdf $h(\theta)$ that describes the unknown parameter. The prior pdf can inject "subjective" information into the inference process because it can be little more than a guess from the user or from some

Osonde Osoba and Bart Kosko are with the Department of Electrical Engineering, Signal and Image Processing Institute, University of Southern California, Los Angeles, California 90089-2564. Sanya Mitaim is with the Department of Electrical and Computer Engineering, Faculty of Engineering, Thammasat University, Pathumthani 12120, Thailand. Contact email: kosko@usc.edu.

consulted expert or other source of authority. Priors can also capture "objective" information from a collateral source of data.

Additive fuzzy systems use if-then rules to map inputs to outputs and thus to model priors or likelihoods. A fuzzy system with enough rules can uniformly approximate any continuous function on a compact domain. Statistical learning algorithms can grow rules from unsupervised clusters in the input-output data or from supervised gradient descent. Fuzzy systems also allow users to add or delete knowledge by simply adding or deleting if-then rules. So they can directly model prior pdfs and approximate them from sample data if it is available. Inverse algorithms can likewise find fuzzy rules that maximize the posterior pdf or functionals based on it. The adaptive fuzzy systems approximate the prior and likelihood pdfs for iterative Bayesian inference and thus differ from the many fuzzified Bayes Theorems in [11], [28] and elsewhere. They preserve the numerical structure of modern Bayesian inference and so also differ from earlier efforts to fuzzify Bayesian inference by using fuzzy-set inputs and other fuzzy contraints [7], [32].

We first demonstrate this fuzzy approximation with the three well-known conjugate priors of Bayesian inference and with a non-conjugate prior. A conjugate prior pdf of one type combines with some randomly sampled data from a likelihood pdf to produce a posterior pdf of the same type: beta priors combine with binomial data to produce beta posteriors, gamma priors combine with Poisson data to produce gamma posteriors, and normal priors combine with normal data to produce normal posteriors. Figures 3-11 below show how adaptive standard-additive-model (SAM) fuzzy systems can approximate these three conjugate priors and their corresponding posteriors. Section II reviews Bayesian inference with these conjugate priors. Section III presents the learning laws that use sample data to tune the fuzzy-system approximators for the six different shaped if-part fuzzy sets in Figure 2. Section IV extends the fuzzy approximation to hierarchical Bayes models where the user puts a second-order prior pdf or a hyperprior on one of the uncertain parameters in the original prior pdf. Section V further extends the fuzzy approach to doubly fuzzy Bayesian inference where separate fuzzy systems approximate the prior and the likelihood. This section also states and proves what we call the Bayesian Approximation Theorem: Uniform fuzzy approximation of the prior and likelihood results in uniform fuzzy approximation of the posterior.

II. Bayesian Statistics and Conjugacy

Bayesian inference models learning as computing a conditional probability based both on new evidence or data and on prior probabilistic beliefs. It builds on the simple Bayes theorem that shows how set-theoretic evidence should update competing prior probabilistic beliefs or hypotheses. The theorem gives the posterior conditional probability $P(H_j|E)$ that the jth hypothesis H_j occurs given that evidence E occurs. The posterior depends on all the converse conditional probabilities $P(E|H_k)$ that E occurs given H_k and on all the unconditional prior probabilities $P(H_k)$ of the disjoint and exhaustive hypotheses $\{H_k\}$:

$$P(H_j|E) = \frac{P(E|H_j)P(H_j)}{P(E)} = \frac{P(E|H_j)P(H_j)}{\sum_k P(E|H_k)P(H_k)}. \quad (1)$$

The result follows from the definition of conditional probability $P(B|A) = P(A \cap B)/P(A)$ for $P(A) > 0$ when the set hypotheses H_j partition the state space of the probability measure P [16], [27].

Bayesian inference or so-called "Bayesian statistics" [1], [8], [9] usually works with a continuous version of (1). Now the parameter value θ corresponds to the hypothesis of interest and the evidence corresponds to the sample values x from a random variable X that depends on θ:

$$f(\theta|x) = \frac{g(x|\theta)h(\theta)}{\int g(x|u)h(u)du} \propto g(x|\theta)h(\theta) \quad (2)$$

where we follow convention and drop the normalizing term that does not depend on θ as we always can if θ has a sufficient statistic [8], [9]. The model (2) assumes that random variable X conditioned on θ admits the random sample $X_1, \ldots, X_n$ with observed realizations $x_1, \ldots, x_n$. So again the posterior pdf $f(\theta|x)$ depends on the converse likelihood $g(x|\theta)$ and on the prior pdf $h(\theta)$. The posterior $f(\theta|x)$ contains the complete Bayesian description of this probabilistic world. Its maximization is a standard optimality criterion in statistical decision making [1], [3], [4], [5], [8], [9].

The Bayes inference structure in (2) involves a radical abstraction. The set or event hypothesis H_j in (1) has become the measurable function or *random variable* Θ that takes on realizations θ according to the prior pdf $h(\theta) : \Theta \sim h(\theta)$. The pdf $h(\theta)$ can make or break the accuracy of the posterior pdf $f(\theta|x)$ because it scales the data pdf $g(x|\theta)$ in (2). The prior itself can come from an expert and thus be "subjective" because it is ultimately an opinion or guess. Or the prior in "empirical Bayes" [3], [8] can come from "objective" data or from statistical hypothesis tests such as chi-squared or Kolmogorov-Smirnov tests for a candidate pdf [9]. Section III shows that the prior can also come from fuzzy rules that in turn come from an expert or from training data or from both.

A. Conjugate Priors

The most common priors tend to be conjugate priors. These priors produce not only closed-form posterior pdfs but posteriors that come from the same family as the prior [1], [4], [8], [26]. The three most common conjugate priors in the literature are the beta, the gamma, and the normal. Table I displays these three conjugacy relationships. The posterior $f(\theta|x)$ is beta if the prior $h(\theta)$ is beta and if the data or likelihood $g(x|\theta)$ is binomial or has a dichotomous Bernoulli structure. The posterior is gamma if the prior is gamma and if the data is Poisson or has a counting structure. The posterior is normal if the prior and data are normal. Conjugate priors permit easy iterative or sequential Bayesian learning because the previous posterior pdf $f_{\text{old}}(\theta|x)$ becomes the new prior pdf $h_{\text{new}}(\theta)$ for the next experiment based on a fresh random sample: $h_{\text{new}}(\theta) = f_{\text{old}}(\theta|x)$. Such conjugacy relations greatly simplify iterative convergence schemes such as Gibbs sampling in Markov chain Monte Carlo estimation of posterior pdfs [3], [8].

TABLE I
CONJUGACY RELATIONSHIPS IN BAYESIAN INFERENCE. A PRIOR PDF OF ONE TYPE COMBINES WITH ITS CONJUGATE LIKELIHOOD PDF TO PRODUCE A POSTERIOR PDF OF THE SAME TYPE.

PRIOR $h(\theta)$	**LIKELIHOOD** $g(x\|\theta)$	**POSTERIOR** $f(\theta\|x)$
Beta	Binomial	Beta$'$
$B(\alpha,\beta)$	bin(n,θ)	$B(\alpha+x,\ \beta+n-x)$
$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\ \theta^{\alpha-1}(1-\theta)^{\beta-1}$	$\binom{n}{x}\theta^x(1-\theta)^{n-x}$	$\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+x)\Gamma(\beta+n-x)}\ \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$
Gamma	Poisson	Gamma$'$
$\Gamma(\alpha,\beta)$	$p(\theta)$	$\Gamma(\alpha+x,\frac{\beta}{1+\beta})$
$\frac{\theta^{\alpha-1}\exp(-\theta/\beta)}{\Gamma(\alpha)\beta^\alpha}$	$e^{-\theta}\ \frac{\theta^x}{x!}$	$\frac{(\theta+\theta\beta)^{\alpha+x}}{\theta\ \Gamma(\alpha+x)\ \beta^{\alpha+x}}\ \exp\left(\frac{-\theta(1+\beta)}{\beta}\right)$
Normal	Normal$'$	Normal$''$
$N(\mu,\tau^2)$	$N(\theta\|\sigma^2)$	$N\left(\frac{\mu\tau^2+x\sigma^2}{\tau^2+\sigma^2},\frac{\tau^2\sigma^2}{\tau^2+\sigma^2}\right)$

1) Beta-Binomial Conjugacy: Consider the beta prior on the unit interval:

$$\Theta\sim\beta(\alpha,\beta):\quad h(\theta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{3}$$

if $0<\theta<1$ for parameters $\alpha>0$ and $\beta>0$. Here Γ is the gamma function $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}dx$. Then Θ has population mean or expectation $E[\Theta]=\alpha/(\alpha+\beta)$. The beta pdf reduces to the uniform pdf if $\alpha=\beta=1$. A beta prior is a natural choice when the unknown parameter θ is the success probability for binomial data such as coin flips or other Bernoulli trials because the beta's support is the unit interval (0, 1) and because the user can adjust the α and β parameters to shape the beta pdf over the interval.

A beta prior is conjugate to binomial data with likelihood pdf $g(x_1,\ldots,x_n|\theta)$. This means that a beta prior $h(\theta)$ combines with binomial sample data to produce a new beta posterior:

$$f(\theta|x)=\frac{\Gamma(n+\alpha+\beta)}{\Gamma(\alpha+x)\Gamma(n+\beta-x)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \tag{4}$$

Here x is the observed sum of n Bernoulli trials and hence is an observed sufficient statistic for θ [9]. So $g(x_1,\ldots,x_n|\theta)=g(x|\theta)$. This beta posterior $f(\theta|x)$ gives the mean-square optimal estimator as the conditional mean $E[\Theta|X=x]=(\alpha+x)/(\alpha+\beta+n)$ if the loss function is squared-error [9]. A beta conjugate relation still holds when negative-binomial or geometric data replaces the binomial data or likelihood. The conjugacy result also extends to the vector case for the Dirichlet or multidimensional beta pdf. A Dirichlet prior is conjugate to multinomial data [4], [24].

2) Gamma-Poisson Conjugacy: Gamma priors are conjugate to Poisson data. The gamma pdf generalizes many right-sided pdfs such as the exponential and chi-square pdfs. The generalized (three-parameter) gamma further generalizes the Weibull and lognormal pdfs. A gamma prior is right-sided and has the form

$$\Theta\sim\gamma(\alpha,\beta):\quad h(\theta)=\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}\qquad\text{if }\theta>0. \tag{5}$$

The gamma random variable Θ has population mean $E[\Theta]=\alpha\beta$ and variance $V[\Theta]=\alpha\beta^2$.

The Poisson sample data $x_1,\ldots,x_n$ comes from the likelihood pdf $g(x_1\ldots,x_n|\theta)=\frac{\theta^{x_1}e^{-\theta}}{x_1!}\cdots\frac{\theta^{x_n}e^{-\theta}}{x_n!}$. The observed Poisson sum $x=x_1+\cdots+x_n$ is an observed sufficient statistic for θ because the Poisson pdf also comes from an exponential family [1], [8]. The gamma prior $h(\theta)$ combines with the Poisson likelihood $g(x|\theta)$ to produce a new gamma posterior $f(\theta|x)$ [9]:

$$f(\theta|x)=\frac{\theta^{(\sum_{k=1}^n x_k+\alpha-1)}e^{-\theta/[\beta/(n\beta+1)]}}{\Gamma(\sum_{k=1}^n x_k+\alpha)[\beta/(n\beta+1)]^{(\sum_{k=1}^n x_k+\alpha)}}. \tag{6}$$

So $E[\Theta|X=x]=(\alpha+x)\beta/(1+\beta)$ and $V[\Theta|X=x]=(\alpha+x)\beta^2/(1+\beta)^2$.

3) Normal-Normal Conjugacy: A normal prior is self-conjugate because a normal prior is conjugate to normal data. A normal prior pdf has the whole real line as its domain and has the form [9]

$$\Theta\sim N(\theta_0,\sigma_0^2):\quad h(\theta)=\frac{1}{\sqrt{2\pi}\sigma_0}e^{-(\theta-\theta_0)^2/2\sigma_0^2} \tag{7}$$

for known population mean θ_0 and known population variance σ_0^2. The normal prior $h(\theta)$ combines with normal sample data from $g(x|\theta)=N(\theta|\sigma^2/n)$ given an observed realization x of the sample-mean sufficient statistic $\overline{X}_n$. This gives the normal posterior pdf $f(\theta|x)=N(\mu_n,\sigma_n^2)$. Here μ_n is the weighted-sum conditional mean $E[\Theta|X=x]=\left(\frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\right)x+\left(\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\right)\theta_0$ and $\sigma_n^2=\left(\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\right)\sigma_0^2$. A hierarchical Bayes model [3], [8] would write any of the these priors as a function of still other random variables and their pdfs as we demonstrate below in Section IV.

III. ADAPTIVE FUZZY APPROXIMATION

Additive fuzzy systems can uniformly approximate continuous functions on compact sets [12], [13], [15]. Hence the set of additive fuzzy systems is dense in the space of such functions. A scalar fuzzy system is the map $F:R^n\to R$ that stores m if-then rules and maps vector inputs x to scalar outputs $F(\theta)$. The prior and likelihood simulations below map not R^n but a compact real interval $[a,\ b]$ into reals. So these systems also satisfy the approximation theorem but at the expense of truncating the domain of pdfs such as the gamma and the normal. Truncation still leaves a proper posterior pdf through the normalization in (2).

A. SAM Fuzzy Systems

A standard additive model (SAM) fuzzy system computes the output $F(\theta)$ by taking the centroid of the sum of the "fired" or scaled then-part sets: $F(\theta)=Centroid(w_1a_1(\theta)B_1+\cdots+w_ma_m(\theta)B_m)$. Then the SAM Theorem states that the output $F(\theta)$ is a simple convex-weighted sum of the then-part set centroids c_j [12], [13], [15], [21]:

$$F(\theta)\ =\ \frac{\sum_{j=1}^m w_ja_j(\theta)V_jc_j}{\sum_{j=1}^m w_ja_j(\theta)V_j}\ =\ \sum_{j=1}^m p_j(\theta)c_j. \tag{8}$$

Here V_j is the finite area of then-part set B_j in the rule "If $X = A_j$ then $Y = B_j$" and c_j is the centroid of B_j. The then-part sets B_j can depend on the input θ and thus their centroids c_j can be functions of θ: $c_j(\theta) = Centroid(B_j(\theta))$. The convex weights $p_1(\theta), \dots, p_m(\theta)$ have the form $p_j(\theta) = \frac{w_j a_j(\theta) V_j}{\sum_{i=1}^m w_i a_i(\theta) V_i}$. The convex coefficients $p_j(\theta)$ change with each input θ. The positive rule weights w_j give the relative importance of the jth rule. They drop out in our case because they are all equal.

The scalar set function $a_j : R \to [0, 1]$ measures the degree to which input $\theta \in R$ belongs to the fuzzy or multivalued set A_j: $a_j(\theta) = Degree(\theta \in A_j)$. The sinc set functions below map into the augmented range $[-.217, 1]$. They require some care in simulations because the denominator in (8) can be zero. We can replace the input θ with θ' in a small neighborhood of θ and so replace the undefined $F(\theta)$ with $F(\theta')$ when the denominator in (8) equals zero. The fuzzy membership value $a_j(\theta)$ "fires" the rule "If $\Theta = A_j$ then $Y = B_j$" in a SAM by scaling the then-part set B_j to give $a_j(\theta)B_j$. The if-part sets can in theory have any shape but in practice they are parametrized pdf-like sets such as those we use below: sinc, Gaussian, triangle, Cauchy, Laplace, and generalized hyperbolic tangent. The if-part sets control the function approximation and involve the most computation in adaptation. Users define a fuzzy system by giving the m corresponding pairs of if-part A_j and then-part B_j fuzzy sets. Many fuzzy systems in practice work with simple then-part fuzzy sets such as congruent triangles or rectangles.

SAMs define "model-free" statistical estimators in the following sense [15], [19], [21]:

$$E[Y|\Theta = \theta] \;=\; F(\theta) \;=\; \sum_{j=1}^{m} p_j(\theta) c_j \tag{9}$$

$$V[Y|\Theta = \theta] = \sum_{j=1}^{m} p_j(\theta)\sigma_{B_j}^2 + \sum_{j=1}^{m} p_j(\theta)[c_j - F(\theta)]^2. \tag{10}$$

The then-part set variance $\sigma_{B_j}^2$ is $\sigma_{B_j}^2 = \int_{-\infty}^{\infty}(y - c_j)^2 p_{B_j}(y)dy$. Then $p_{B_j}(y) = b_j(y)/V_j$ is an integrable pdf if $b_j : R \to [0,\ 1]$ is the integrable set function of then-part set B_j. The conditional variance $V[Y|\Theta = \theta]$ gives a direct measure of the uncertainty in the SAM output $F(\theta)$ based on the inherent uncertainty in the stored then-part rules. This defines a type of confidence surface for the fuzzy system [19]. The first term in the conditional variance (10) measures the inherent uncertainty in the then-part sets given the current rule firings. The second term is an interpolation penalty because the rule "patches" $A_j \times B_j$ cover different regions of the input-output product space. The shape of the then-part sets affects the conditional variance of the fuzzy system but affects the output $F(\theta)$ only to the extent that the then-part sets B_j have different centroids c_j or areas V_j. The adaptive function approximations below tune only these two parameters of each then-part set. The conditional mean (9) and variance (10) depend on the realization $\Theta = \theta$ and so generalize the corresponding unconditional mean and variance of mixture densities [8].

A SAM fuzzy system F can always approximate a function

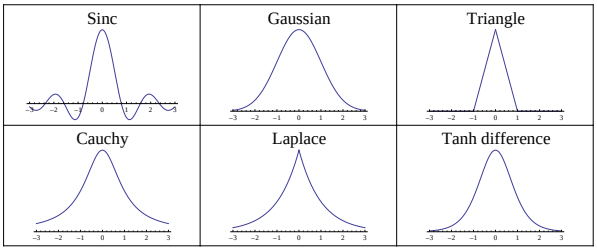

Fig. 2. Six types of if-part fuzzy sets in conjugate prior approximations. Each type of set produces its own adaptive SAM learning law for tuning its location and dispersion parameters: (a) sinc set, (b) Gaussian set, (c) triangle set, (d) Cauchy set, (e) Laplace set, and (f) a generalized hyperbolic-tangent set. The sinc shape performed best in most approximations of conjugate priors and the corresponding fuzzy-based posteriors.

f or $F \approx f$ if the fuzzy system contains enough rules. But multidimensional fuzzy systems $F : R^n \to R$ suffer exponential rule explosion in general because they require $\mathcal{O}(k^n)$ rules [10], [14], [22]. Optimal rules tend to reside at the extrema or turning points of the approximand f and so optimal fuzzy rules "patch the bumps" [14]. Learning tends to quickly move rules to these extrema and to fill in with extra rules between the extremum-covering rules. The supervised learning algorithms can involve extensive computation in higher dimensions [20], [21]. Our fuzzy prior approximation $F : R \to R$ maps scalars to scalars so it requires only $\mathcal{O}(k)$ rules and thus does not suffer rule explosion. But Theorem 3 below shows that iterative Bayesian inference can produce its own rule explosion.

B. The Watkins Representation Theorem

Fuzzy systems can exactly represent a bounded pdf with a known closed form. Watkins has shown that in many cases a SAM system F can exactly represent a function f in the sense that $F = f$ [29], [30]. The Watkins Representation Theorem states that $F = f$ if f is bounded and if we know the closed form of f. The result is stronger than this because the SAM system F exactly represents f with just *two* rules with equal weights $w_1 = w_2$ and equal then-part set volumes $V_1 = V_2$:

$$F(\theta) \;=\; \frac{\sum_{j=1}^{2} w_j a_j(\theta) V_j c_j}{\sum_{j=1}^{2} w_j a_j(\theta) V_j} \tag{11}$$

$$\;=\; \frac{a(\theta)c_1 + a^c(\theta)c_2}{a(\theta) + a^c(\theta)} \tag{12}$$

$$\;=\; f(\theta) \tag{13}$$

if $a_1(\theta) = a(\theta) = \dfrac{\sup f - f(\theta)}{\sup f - \inf f}$, $a_2(\theta) = a^c(\theta) = 1 - a(\theta)$, $c_1 = \inf f$, and $c_2 = \sup f$.

The representation technique builds f directly into the structure of the two if-then rules. Let $h(\theta)$ be any bounded prior pdf such as the $\beta(8, 5)$ pdf in the simulations below. Then $F(\theta) = h(\theta)$ holds for all realizations θ if the SAM's two rules have the form "If $\Theta = A$ then $Y = B_1$" and "If $\Theta =$ not-A then $Y = B_2$" for the if-part set function

$$a(\theta) = \frac{\sup h - h(\theta)}{\sup h - \inf h} = 1 - \frac{11^{11}}{7^7 4^4}\theta^7(1 - \theta)^4 \tag{14}$$

if $\Theta \sim \beta(8, 5)$. Then-part sets B_1 and B_2 can have any shape from rectangles to Gaussians so long as $0 < V_1 = V_2 < \infty$ with centroids $c_1 = \inf h = 0$ and $c_2 =$

$\sup h = \frac{\Gamma(13)}{\Gamma(8)\Gamma(5)}(\frac{7}{11})^7(\frac{4}{11})^4$. So the Watkins Representation Theorem lets a SAM fuzzy system directly absorb a closed-form bounded prior $h(\theta)$ if it is available. The same holds for a bounded likelihood or posterior pdf.

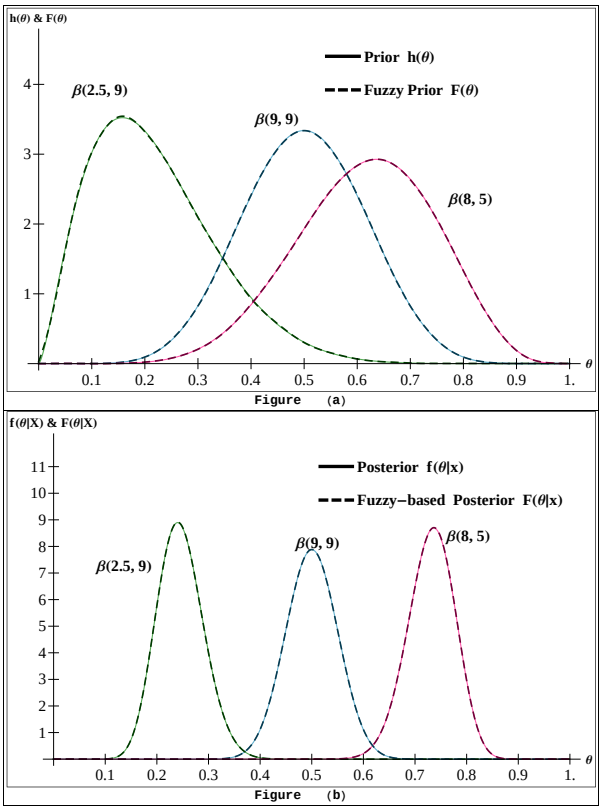

Fig. 3. Comparison of conjugate beta priors and posteriors with their fuzzy approximators. (a) an adapted sinc-SAM fuzzy system $F(\theta)$ with 15 rules approximates the three conjugate beta priors $h(\theta)$: $\beta(2.5, 9)$, $\beta(9, 9)$, and $\beta(8, 5)$. (b) the sinc-SAM fuzzy priors $F(\theta)$ in (a) produce the SAM-based approximators $F(\theta|x)$ of the three corresponding beta posteriors $f(\theta|x)$ for the three corresponding binomial likelihood pdfs $g(x|\theta)$ with $n = 80$: $bin(20, 80)$, $bin(40, 80)$, and $bin(60, 80)$ where $g(x|\theta) = bin(x, 80) = \frac{80!}{x!(80-x)!}\theta^x(1-\theta)^{80-x}$. So $X \sim bin(x, 80)$ and $X = 20$ mean that there were 20 successes out of 80 trials in an experiment where the probability of success was θ. Each of the three fuzzy approximations cycled 6,000 times through 500 uniform training samples from the corresponding beta priors.

C. ASAM Learning Laws

An adaptive SAM (ASAM) F can quickly approximate a prior $h(\theta)$ (or likelihood) if the following supervised learning laws have access to adequate samples $h(\theta_1), h(\theta_2), \dots$ from the prior. This may mean in practice that the ASAM trains on the same numerical data that a user would use to conduct a chi-squared or Kolmogorov-Smirnov hypothesis test for a candidate pdf. Figure 4 shows that an ASAM can learn the prior pdf even from noisy random samples drawn from the pdf. Unsupervised clustering techniques can also train an ASAM if there is sufficient cluster data [12], [15], [31]. The ASAM prior simulations in the next section show how F approximates $h(\theta)$ when the ASAM trains on random samples from the prior. These approximations bolster the case that ASAMs will in practice learn the appropriate prior that corresponds to the available collateral data.

ASAM supervised learning uses gradient descent to tune the parameters of the set functions a_j as well as the then-part areas V_j (and weights w_j) and centroids c_j. The learning laws follow from the SAM's convex-sum structure (8) and the chain-rule decomposition $\frac{\partial E}{\partial m_j} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial a_j}\frac{\partial a_j}{\partial m_j}$ for SAM parameter m_j and error E in the generic gradient-descent algorithm [15], [21]

$$m_j(t+1) = m_j(t) - \mu_t \frac{\partial E}{\partial m_j} \tag{15}$$

where μ_t is a learning rate at iteration t. We seek to minimize the squared error

$$E(\theta) = \frac{1}{2}(f(\theta) - F(\theta))^2 = \frac{1}{2}\varepsilon(\theta)^2 \tag{16}$$

of the function approximation. Let m_j denote any parameter in the set function a_j. Then the chain rule gives the gradient of the error function with respect to the respective if-part set parameter m_j, the centroid c_j, and the volume V_j:

$$\frac{\partial E}{\partial m_j} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial a_j}\frac{\partial a_j}{\partial m_j} \tag{17}$$

$$\frac{\partial E}{\partial c_j} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial c_j} \tag{18}$$

$$\frac{\partial E}{\partial V_j} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial V_j} \tag{19}$$

with partial derivatives [15], [21]

$$\frac{\partial E}{\partial F} = -(f(\theta) - F(\theta)) = -\varepsilon(\theta) \tag{20}$$

$$\frac{\partial F}{\partial a_j} = [c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}. \tag{21}$$

The SAM ratio (8) with equal rule weights $w_1 = \cdots = w_m$ gives [15], [21]

$$\frac{\partial F}{\partial c_j} = \frac{a_j(\theta)V_j}{\sum_{i=1}^m a_i(\theta)V_i} = p_j(\theta) \tag{22}$$

$$\frac{\partial F}{\partial V_j} = \frac{a_j(\theta)[c_j - F(\theta)]}{\sum_{i=1}^m a_i(\theta)V_i} = [c_j - F(\theta)]\frac{p_j(\theta)}{V_j}. \tag{23}$$

Then the learning laws for the then-part set centroids c_j and volume V_j have the final form

$$c_j(t+1) = c_j(t) + \mu_t\varepsilon(\theta)p_j(\theta) \tag{24}$$

$$V_j(t+1) = V_j(t) + \mu_t\varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{V_j}. \tag{25}$$

The learning laws for the if-part set parameters follow in like manner by expanding $\frac{\partial a_j}{\partial m_j}$ in (17).

The simulations below tune the location m_j and dispersion d_j parameters of the if-part set functions a_j for sinc, Gaussian, triangle, Cauchy, Laplace, and generalized hyperbolic tangent if-part sets. Figure 2 shows an example of each of these six fuzzy sets with the following learning laws.

1) Sinc ASAM learning law: The sinc set function a_j has the form

$$a_j(\theta) = \sin\left(\frac{\theta - m_j}{d_j}\right) \Big/ \left(\frac{\theta - m_j}{d_j}\right) \tag{26}$$

with parameter learning laws [15], [21]

$$m_j(t+1) = m_j(t) + \mu_t\varepsilon(\theta)[c_j - F(\theta)]\times \frac{p_j(\theta)}{a_j(\theta)}\left(a_j(\theta) - \cos\left(\frac{\theta - m_j}{d_j}\right)\right)\frac{1}{\theta - m_j} \tag{27}$$

$$d_j(t+1) = d_j(t) + \mu_t\varepsilon(\theta)[c_j - F(\theta)]\times \frac{p_j(\theta)}{a_j(\theta)}\left(a_j(\theta) - \cos\left(\frac{\theta - m_j}{d_j}\right)\right)\frac{1}{d_j}. \tag{28}$$

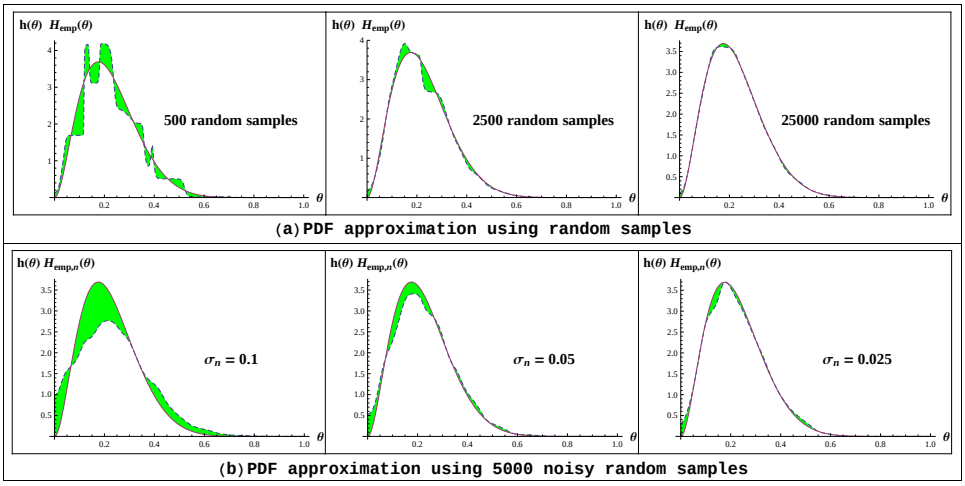

Fig. 4. ASAMs can use a limited number of random samples or noisy random samples to estimate the sampling pdf. The ASAMs for these examples use the tanh set function with 15 rules and they run for 6000 iterations. The ASAMs approximate empirical pdfs from the different sets of random samples. The shaded regions represent the approximation error between the ASAM estimate and the sampling pdf. Part (a) compares the $\beta(3, 10.4)$ pdf with ASAM approximations for some $\beta(3, 10.4)$ empirical pdfs. Each empirical pdf is a scaled histogram for a set of N random samples. The figure shows comparisons for the cases $N = 500$, 2500, 25000. Part (b) compares the $\beta(3, 10.4)$ pdf with ASAM approximations of 3 $\beta(3, 10.4)$ random sample sets corrupted by independent noise. Each set has 5000 random samples. The noise is zero-mean additive white Gaussian noise. The standard deviations σ_n of the additive noise are 0.1, 0.05, and 0.025. The plots show that the ASAM estimate gets better as the number of samples increases. The ASAM has difficulty estimating tail probabilities when the additive noise variance gets large.

2) Gaussian ASAM learning law: The Gaussian set function a_j has the form

$$a_j(\theta) = \exp\left\{-\left(\frac{\theta - m_j}{d_j}\right)^2\right\} \tag{29}$$

with parameter learning laws

$$m_j(t+1) = m_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)]\frac{\theta - m_j}{d_j^2} \tag{30}$$

$$d_j(t+1) = d_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)]\frac{(\theta - m_j)^2}{d_j^3}. \tag{31}$$

3) Triangle ASAM learning law: The triangle set function has the form

$$a_j(\theta) = \begin{cases} 1 - \frac{m_j - \theta}{l_j} & \text{if } m_j - l_j \leq \theta \leq m_j \\ 1 - \frac{\theta - m_j}{r_j} & \text{if } m_j \leq \theta \leq m_j + r_j \\ 0 & \text{else} \end{cases} \tag{32}$$

with parameter learning laws

$$m_j(t+1) = \begin{cases} m_j(t) - \mu_t \varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}\frac{1}{l_j} \\ \qquad \text{if } m_j - l_j < \theta < m_j \\ m_j(t) + \mu_t \varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}\frac{1}{r_j} \\ \qquad \text{if } m_j < \theta < m_j + r_j \\ m_j(t) \quad \text{else} \end{cases} \tag{33}$$

$$l_j(t+1) = \begin{cases} l_j(t) + \mu_t \varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}\frac{m_j - \theta}{l_j^2} \\ \qquad \text{if } m_j - l_j < \theta < m_j \\ l_j(t) \quad \text{else} \end{cases} \tag{34}$$

$$r_j(t+1) = \begin{cases} r_j(t) + \mu_t \varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}\frac{\theta - m_j}{r_j^2} \\ \qquad \text{if } m_j < \theta < m_j + r_j \\ r_j(t) \quad \text{else} \end{cases} \tag{35}$$

The Gaussian learning laws (30)-(31) can approximate the learning laws for the symmetric triangle set function $a_j(\theta) = \max\{0, 1 - \frac{|\theta - m_j|}{d_j}\}$.

4) Cauchy ASAM learning law: The Cauchy set function a_j has the form

$$a_j(\theta) = \frac{1}{1 + \left(\frac{\theta - m_j}{d_j}\right)^2} \tag{36}$$

with parameter learning laws

$$m_j(t+1) = m_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)]\frac{\theta - m_j}{d_j^2} a_j(\theta) \tag{37}$$

$$d_j(t+1) = d_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)]\frac{(\theta - m_j)^2}{d_j^3} a_j(\theta). \tag{38}$$

5) Laplace ASAM learning law: The Laplace or double-exponential set function a_j has the form

$$a_j(\theta) = \exp\left\{-\frac{|\theta - m_j|}{d_j}\right\} \tag{39}$$

with parameter learning laws

$$m_j(t+1) = m_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)]\text{sign}(\theta - m_j)\frac{1}{d_j} \tag{40}$$

$$d_j(t+1) = d_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)]\text{sign}(\theta - m_j)\frac{|\theta - m_j|}{d_j^2} \tag{41}$$

6) Generalized hyperbolic tangent ASAM learning law: The generalized hyperbolic tangent set function has the form

$$a_j(\theta) = 1 + \tanh\left(-\left(\frac{\theta - m_j}{d_j}\right)^2\right) \tag{42}$$

with parameter learning laws

$$m_j(t+1) = m_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)](2 - a_j(\theta))\frac{\theta - m_j}{d_j^2} \tag{43}$$

$$d_j(t+1) = d_j(t) + \mu_t \varepsilon(\theta) p_j(\theta)[c_j - F(\theta)](2 - a_j(\theta))\frac{(\theta - m_j)^2}{d_j^3} \tag{44}$$

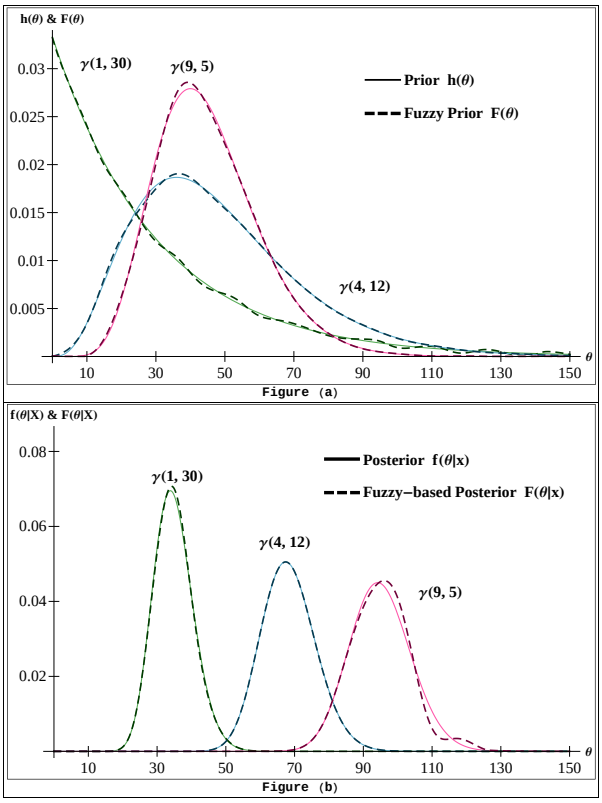

Fig. 5. Comparison of conjugate gamma priors and posteriors with their fuzzy approximators. (a) an adapted sinc-SAM fuzzy system $F(\theta)$ with 15 rules approximates the three conjugate gamma priors $h(\theta)$: $\gamma(1, 30)$, $\gamma(4, 12)$, and $\gamma(9, 5)$. (b) the sinc-SAM fuzzy priors $F(\theta)$ in (a) produce the SAM-based approximators $F(\theta|x)$ of the three corresponding gamma posteriors $f(\theta|x)$ for the three corresponding Poisson likelihood pdfs $g(x|\theta)$: $p(35)$, $p(70)$, and $p(105)$ where $g(x|\theta) = p(x) = \theta^x e^{-\theta}/x!$. Each of the three fuzzy approximations cycled 6,000 times through 1,125 uniform training samples from the corresponding gamma priors.

We can also reverse the learning process and adapt the SAM if-part and then-part set parameters by maximizing a given closed-form posterior pdf $f(\theta|x)$. The basic Bayesian relation (2) above leads to the following application of the chain rule for a set parameter m_j:

$$\frac{\partial f(\theta|x)}{\partial m_j} \propto g(x|\theta)\frac{\partial F}{\partial m_j} \tag{45}$$

since $\frac{\partial g}{\partial F} = 0$ because the likelihood $g(x|\theta)$ does not depend on the fuzzy system F. The chain rule gives $\frac{\partial F}{\partial m_j} = \frac{\partial F}{\partial a_j}\frac{\partial a_j}{\partial m_j}$ and similarly for the other SAM parameters. Then the above learning laws can eliminate the product of partial derivatives to produce a stochastic gradient ascent or maximum-a-posteriori or MAP learning law for the SAM parameters.

D. ASAM Approximation Simulations

We simulated six different types of adaptive SAM fuzzy systems to approximate the three standard conjugate prior pdfs and their corresponding posterior pdfs. The six types of ASAMs corresponded to the six if-part sets in Figure 2 and their learning laws above. We combined C++ software for the ASAM approximations with Mathematica to compute the fuzzy-based posterior $F(\theta|x)$ using (2). Mathematica's NIntegrate program computed the mean-squared errors between the conjugate prior $h(\theta)$ and the fuzzy-based prior $F(\theta)$ and between the posterior $f(\theta|x)$ and the fuzzy posterior $F(\theta|x)$.

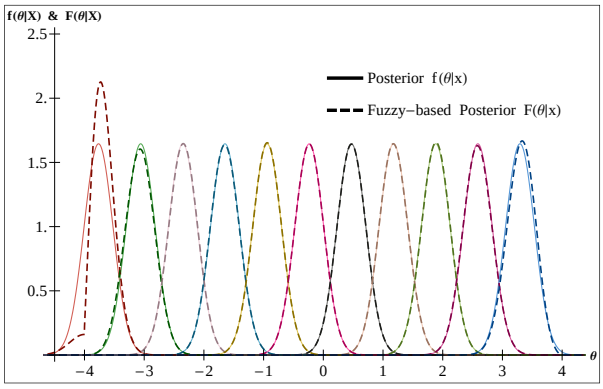

Fig. 6. Comparison of 11 conjugate normal posteriors with their fuzzy-based approximators based on a standard normal prior and 11 different normal likelihoods. An adapted sinc-SAM approximator with 15 rules first approximates the standard normal prior $h(\theta) = N(0, 1)$ and then combines with the likelihood pdf $g(x|\theta) = N(\theta|\frac{1}{16})$. The variance is $1/16$ because x is the observed sample mean of 16 standard-normal random samples $X_k \sim N(0, 1)$. The 11 priors correspond to the 11 likelihoods $g(x|\theta)$ with $x = -4$, -3.25, -2.5, -1.75, -1, -0.25, 0.5, 1.25, 2, 2.75, and 3.5. The fuzzy approximation cycled 6,000 times through 500 uniform training samples from the standard-normal prior.

Each ASAM simulation used uniform samples from a prior pdf $h(\theta)$. The program evenly spaced the initial if-part sets and assigned them equal but experimental dispersion values. The initial then-part sets had unit areas or volumes. The initial then-part centroids corresponded to the prior pdf's value at the location parameters of the if-part sets. A single learning iteration began with computing the approximation error at each uniformly spaced sample point. The program cycled through all rules for each sample value and then updated each rule's if-part and then-part parameters according to the appropriate ASAM learning law. Each adapted parameter had a harmonic-decay learning rate $\mu_t = \frac{c}{t}$ for learning iteration t. Experimentation picked the numerator constants c for the various parameters.

The approximation figures show representative simulation results. Figure 1 used Cauchy if-part sets for illustration only and not because they gave a smaller mean-squared error than sinc sets did. Figures 3-6 used sinc if-part sets even though we simulated all six types of if-part sets for all three types of conjugate priors. Simulations demonstrated that all 6 set functions produce good approximations for the prior pdfs. The sinc ASAM usually performed best. We truncated the gamma priors at the right-side value of 150 and truncated the normal priors at -4 and 4 because the overlap between the truncated prior tails and the likelihood pdfs $g(x|\theta)$ were small. The likelihood functions $g(x|\theta)$ had narrow dispersions relative to the truncated supports of the priors. Larger truncation values or appended fall-off tails can accommodate unlikely x values in other settings. We also assumed that the priors were strictly positive. So we bounded the ASAM priors to a small positive value $(F(\theta) \geq 10^{-3})$ to keep the denominator integral in (2) well-behaved.

Figure 1 used only one fuzzy approximation. The fuzzy approximation of the $\beta(8, 5)$ prior had mean-squared error 4.2×10^{-4}. The Cauchy-ASAM learning algorithm used 500 uniform samples for 6,000 iterations.

The fuzzy approximation of the beta priors $\beta(2.5, 9)$, $\beta(9, 9)$, and $\beta(8, 5)$ in Figure 3 had respective mean-squared

errors 1.3×10^{-4}, 2.3×10^{-5}, and 1.4×10^{-5}. The sinc-ASAM learning used 500 uniform samples from the unit interval for 6,000 training iterations. The corresponding conjugate beta posterior approximations had respective mean-squared errors 3.0×10^{-5}, 6.9×10^{-6}, and 3.8×10^{-5}.

The fuzzy approximation of the gamma priors $\gamma(1, 30)$, $\gamma(4, 12)$, and $\gamma(9, 5)$ in Figure 5 had respective mean-squared errors 5.5×10^{-5}, 3.6×10^{-6}, and 7.9×10^{-6}. The sinc-ASAM learning used 1,125 uniform samples from the truncated interval $[0, 150]$ for 6,000 training iterations. The corresponding conjugate gamma posterior approximations had mean-squared errors 2.3×10^{-5}, 2.1×10^{-7}, and 2.3×10^{-4}.

The fuzzy approximation of the single standard-normal prior that underlies Figure 6 had mean-squared error of 7.7×10^{-6}. The sinc-ASAM learning used 500 uniform samples from the truncated interval $[-4, 4]$ for 6,000 training iterations. Table II gives the MSEs for the normal posteriors.

Sample Mean	MSE	Sample Mean	MSE
-4	0.12		
-3.25	1.9×10^{-3}	0.5	1.1×10^{-5}
-2.5	3×10^{-4}	1.25	6.5×10^{-5}
-1.75	1.5×10^{-4}	2	1.6×10^{-4}
-1	3.1×10^{-5}	2.75	3×10^{-4}
-0.25	2.2×10^{-6}	3.5	7.6×10^{-3}

TABLE II
MEAN SQUARED ERRORS FOR THE 11 NORMAL POSTERIOR APPROXIMATIONS

The generalized-hyperbolic-tanh ASAMs in Figure 4 learn the beta prior $\beta(3, 10.4)$ from both noiseless and noisy random-sample (i.i.d.) $x_1, x_2, \ldots$ draws from the "unknown" prior because the ASAMs use only the histogram or empirical distribution of the pdf. The Glivenko-Cantelli Theorem [2] ensures that the empirical distribution converges uniformly to the original distribution. So sampling from the histogram of random samples increasingly resembles sampling directly from the unknown underlying pdf as the sample size increases. This ASAM learning is robust in the sense that the fuzzy systems still learn the pdf if independent white noise corrupts the random-sample draws.

The simulation draws N random samples $x_1, x_2, \ldots, x_N$ from the pdf $h(\theta) = \beta(3, 10.40)$ and then bins them into 50 equally spaced bins of length $\Delta\theta = 0.02$. We generate an empirical pdf $h_{emp}(\theta)$ for the beta distribution by rescaling the histogram. The rescaling converts the histogram into a staircase approximation of the pdf $h(\theta)$:

$$h_{emp}(\theta) \quad = \quad \sum_{m=1}^{\#\text{ of bins}} \frac{p[m] rect(\theta - \theta_b[m])}{N\Delta\theta} \qquad (46)$$

where $p[m]$ is the number of random samples in bin m and where $\theta_b[m]$ is the central location of the m^{th} bin. The ASAM generates an approximation $H_{emp}(\theta)$ for the empirical distribution $h_{emp}(\theta)$. Figure 4(a) shows comparisons between $H_{emp}(\theta)$ and $h(\theta)$.

The second example starts with $5,000$ random samples of the $\beta(3, 10.4)$ distribution. We add zero-mean white Gaussian noise to the random samples. The noise is independent of the random samples. The examples use respective noise standard deviations of 0.1, 0.05, and 0.025 in the three separate cases. The ASAM produces an approximation $H_{emp,n}(\theta)$ for this noise-modified function $h_{emp,n}(\theta)$. Figure 4(b) shows comparisons between $H_{emp,n}(\theta)$ to $h(\theta)$. The approximands h_{emp} and $h_{emp,n}$ in Figures 4 (a) and (b) are random functions. So these functions and their ASAM approximators are sample cases.

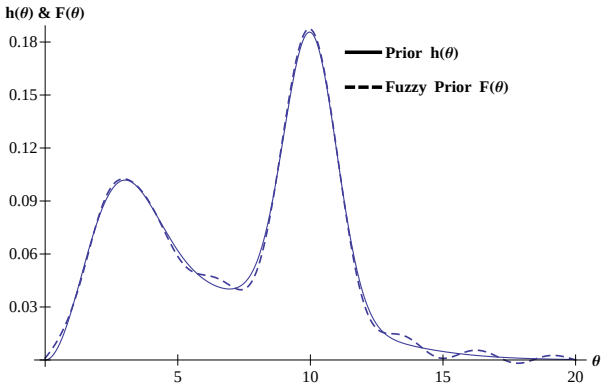

Fig. 7. Comparison of a non-conjugate prior pdf $h(\theta)$ and its fuzzy approximator $H(\theta)$. The pdf $h(\theta)$ is a convex mixture of normal and Maxwell pdfs: $h(\theta) = 0.4N(10, 1) + 0.3M(2) + 0.3M(5)$. The Maxwell pdf $M(\sigma)$ is $\theta^2 e^{-\theta^2/2\sigma^2}$ for $\theta \geq 0$ and 0 for $\theta \leq 0$. An adaptive sinc-SAM generated $H(\theta)$ using 15 rules and 6000 training iterations on 500 uniform samples of the $h(\theta)$.

E. Non-conjugate Priors

The ASAM technique can also approximate non-conjugate priors and their corresponding posteriors. We defined a prior pdf $h(\theta)$ as a convex bimodal mixture of normal and Maxwell pdfs: $h(\theta) = 0.4N(10, 1)+0.3M(2)+0.3M(5)$. The Maxwell pdfs have the form

$$\theta \sim M(\sigma): \quad h(\theta) = \theta^2 e^{-\frac{\theta^2}{2\sigma^2}} \qquad \text{if} \quad \theta > 0. \qquad (47)$$

The prior pdf modeled a location parameter for the normal mixture likelihood function: $g(x|\theta) = 0.7N(\theta, 2.25) + 0.3N(\theta + 8, 1)$. The prior $h(\theta)$ is not conjugate with respect to this likelihood function $g(x|\theta)$. Figures 7 and 8 show the ASAM approximations of the respective prior and posterior.

The ASAM used sinc set functions to generate a fuzzy approximator $H(\theta)$ for the prior $h(\theta)$. The ASAM used 15 rules and 6000 iterations on 500 uniform samples of $h(\theta)$. Figures 7 and 8 show the quality of the prior and posterior fuzzy approximators. This example shows that fuzzy Bayesian approximation still works for non-conjugate pdfs.

F. Closed-Form SAM Posterior Estimates

The next theorem shows that the SAM's convex-weighted-sum structure passes over into the structure of the fuzzy-based posterior $F(\theta|x)$. The result is a *generalized* SAM [15] because the then-part centroids c_j are no longer constant but vary both with the observed data x and the parameter value θ. This simplified structure for the posterior $F(\theta|x)$ comes at the expense in general of variable centroids that require several integrations for each observation x.

Theorem 1: The fuzzy posterior approximator is a SAM:

$$F(\theta|x) \quad = \quad \sum_{j=1}^{m} p_j(\theta) c'_j(x|\theta) \qquad (48)$$

where the generalized then-part set centroids $c'_j(x|\theta)$ have the form

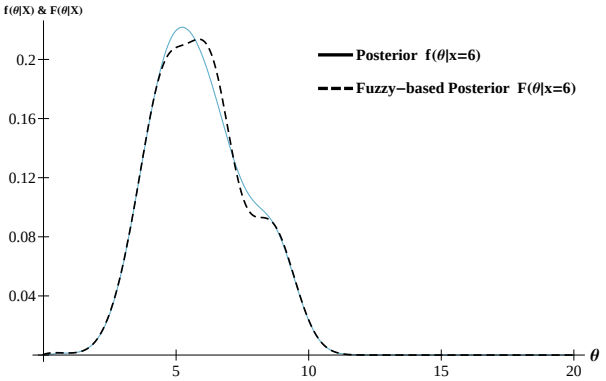

Fig. 8. Approximation of non-conjugate posterior pdf. Comparison of a non-conjugate posterior pdf $f(\theta|x)$ and its fuzzy approximator $F(\theta|x)$. The fuzzy prior $H(\theta)$ and the mixture likelihood function $g(x|\theta) = 0.7N(\theta|2.25) + 0.3N(\theta+8,1)$ produce the fuzzy approximator of the posterior pdf $F(\theta|x)$. The figure shows $F(\theta|x)$ for the single observation $x = 6$.

$$c_j'(\theta|x) = \frac{c_j(\theta)g(x|\theta)}{\sum_{i=1}^{m}\int_{\mathcal{D}} g(x|u)p_i(u)c_i(u)\,du} \tag{49}$$

for sample space $\mathcal{D}$.

We next state two corollaries that hold in special cases that avoid the integration in (49) and thus are computationally tractable.

Corollary 1.1: Suppose $g(x|\theta)$ approximates a Dirac delta function centered at x: $g(x|\theta) \approx \delta(\theta - x)$. Then $c_j'(\theta|x)$ in (49) becomes

$$c_j'(\theta|x) \;\approx\; \frac{c_j g(x|\theta)}{F(x)}. \tag{50}$$

This special case arises when $g(x|\theta)$ concentrates on a region $\mathcal{D}_g \subset \mathcal{D}$ if $\mathcal{D}_g$ is much smaller than $\mathcal{D}_{p_j} \subset \mathcal{D}$ and if $p_j(\theta)$ concentrates on $\mathcal{D}_{p_j}$.

So a learning law for $F(\theta|x)$ needs to update only each then-part centroid c_j by scaling it with $g(x|\theta)/F(x)$ for each observation x. This involves a substantially lighter computation than does the integration in (49).

The delta-pulse approximation $g(x|\theta) \approx \delta(\theta - x)$ holds for narrow bell curves such as normal or Cauchy pdfs when their variance or dispersion is small. It holds in the limit as the equality $g(x|\theta) = \delta(\theta - x)$ in the much more general case of alpha-stable pdfs [17], [25] with any shape if x is the location parameter of the stable pdf and if the dispersion γ goes to zero. Then the characteristic function is the complex exponential $e^{ix\omega}$ and thus Fourier transformation gives the pdf $g(x|\theta)$ exactly as the Dirac delta function [18]: $\lim_{\gamma \to 0} g(x|\theta) = \delta(\theta - x)$. Then

$$F(\theta|x) \;=\; \sum_{j=1}^{m} p_j(\theta)\left(\frac{c_j g(x|\theta)}{F(x)}\right) \tag{51}$$

The approximation fails for a narrow binomial $g(x|\theta)$ unless scaling maintains unity status for the mass of $g(x|\theta)$ in (78) for a given n.

Corollary 1.2: Suppose we can approximate the likelihood $g(x|\theta)$ with constant $g(x|m_j)$ and then-part set centroids $c_j(\theta)$ with constant $c_j(m_j)$ over $\mathcal{D}_{p_j}$. Then $c_j'(\theta|x)$ in (49) becomes

$$c_j'(\theta|x) \;\approx\; \frac{c_j(\theta)g(x|\theta)}{\sum_{i=1}^{m} g(x|m_i)U_{p_i}c_i(m_j)} \tag{52}$$

where $U_{p_j} = \int_{\mathcal{D}_{p_j}} p_j(u)du$.

We can pre-compute or estimate the if-part volume U_{p_j} in advance. So (52) also gives a generalized SAM structure and another tractable way to adapt the variable then-part centroids $c_j'(x|\theta)$.

This second special case holds for the normal likelihood pdf $g(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma_0}e^{-(x-\theta)^2/2\sigma_0^2}$ if the widths or dispersions d_j of the if-part sets are small compared with σ_0 and if there are a large number m of fuzzy if-then rules that jointly cover $\mathcal{D}_g$. This occurs if $D_g = (\theta - 3\sigma_0, \theta + 3\sigma_0)$ with if-part dispersions $d_j = \sigma_0/m$ and locations m_j. Then $p_j(\theta)$ concentrates on some $\mathcal{D}_{p_j} = (m_j - \epsilon, m_j + \epsilon)$ where $0 < \epsilon \ll \sigma_0$ and so $p_j(\theta) \approx 0$ for $\theta \notin \mathcal{D}_{p_j}$. Then $\frac{x-m_j \pm \epsilon}{\sigma_0} \approx \frac{x-m_j}{\sigma_0}$ since $\epsilon \ll \sigma_0$. So $\frac{x-\theta}{\sigma_0} \approx \frac{x-m_j}{\sigma_0}$ for all $\theta \in \mathcal{D}_{p_j}$ and thus

$$g(x|\theta) \approx \frac{1}{\sqrt{2\pi}\sigma_0}e^{-(x-m_j)^2/2\sigma_0^2} = g(x|m_j) \tag{53}$$

for $\theta \in \mathcal{D}_{p_j}$. Then (83) holds.

This special case also holds for the binomial $g(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, \ldots, n$ if $n \ll m$ and thus if there are fewer Bernoulli trials n than fuzzy if-then rules m in the SAM system. It holds because $g(x|\theta)$ concentrates on $\mathcal{D}_g$ and because $\mathcal{D}_g$ is wide compared with $\mathcal{D}_{p_j}$ when $m \gg n$. This case also holds for the Poisson $g(x|\theta) = \frac{1}{x!}\theta^x e^{-\theta}$ if the number of times x that a discrete event occurs is small compared with the number m of SAM rules that jointly cover $\mathcal{D}_g = (\frac{x}{2}, \frac{3x}{2})$ because again $\mathcal{D}_g$ is large compared with $\mathcal{D}_{p_j}$. So (83) follows.

IV. Fuzzy Hierarchical Bayesian Inference

Adaptive fuzzy approximation can also apply to second-order priors or so-called *hierarchical Bayes* techniques [3], [8]. Here the user puts a new prior or *hyperprior* pdf on an uncertain parameter that appears in the original prior pdf. This new hyperprior pdf can itself have a random parameter that leads to yet another new prior or hyper-hyperprior pdf and so on up the hierarchy of prior models. We will demonstrate the hierarchical technique in the common case where an inverse-gamma hyperprior pdf models the uncertainty in the unknown variance of a normal prior pdf. This is the scalar case of the conjugate inverse Wishart prior [3] that often models the uncertainty in the covariance matrix of a normal random vector.

Suppose again that the posterior pdf $f(\theta|x)$ is approximately the product of the likelihood pdf $g(x|\theta)$ and the prior pdf $h(\theta)$:

$$f(\theta|x) \;\sim\; g(x|\theta)\,h(\theta). \tag{54}$$

But now suppose that the prior pdf $h(\theta)$ depends on an uncertain parameter $\tau : h(\theta|\tau)$. We will model the uncertainty involving τ by making τ a random variable T with its own pdf or hyperprior pdf $\pi(\tau)$. Conditioning the original prior $h(\theta)$ on τ adds a new dimension to the posterior pdf:

$$f(\theta|\tau|x) \;\sim\; g(x|\theta)\,h(\theta|\tau)\,\pi(\tau). \tag{55}$$

But marginalizing or integrating over τ removes this extra dimension and restores the original posterior pdf:

$$f(\theta|x) \;\sim\; \int g(x|\theta)\,h(\theta|\tau)\,\pi(\tau)\,d\tau. \tag{56}$$

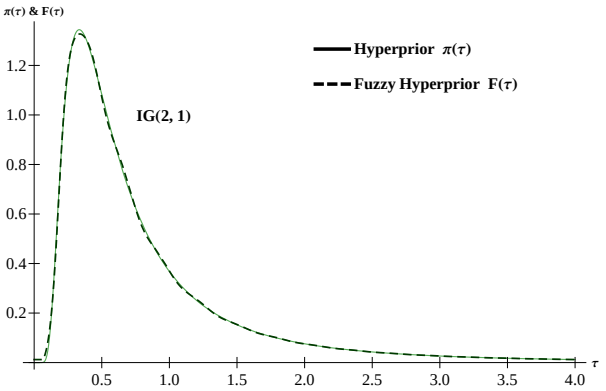

Fig. 9. Comparison of inverse-gamma (IG) hyperprior $\pi(\tau)$ and its fuzzy approximation. The hyperprior pdf is the $IG(2,1)$ pdf that describes the random parameter τ that appears as the variance in a normal prior. The approximating fuzzy hyperprior $F(\tau)$ used 15 rules in a SAM with Gaussian if-part sets. The fuzzy approximator used 1000 uniform samples from $[0,4]$ and 6000 training iterations.

Thus hierarchical Bayes has the benefit of working with a more flexible and descriptive prior but at the computational cost of a new integration. The approach of empirical Bayes [3], [8] would simply replace the random variable τ with a numerical proxy such as its most probable value. That approach is simpler to compute but ignores most of the information in the hyperprior pdf.

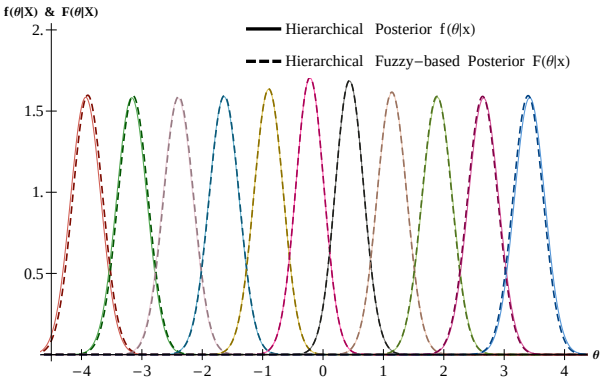

Fig. 10. Hierarchical Bayes posterior pdf approximation using a fuzzy hyperprior. The plot shows the fuzzy approximation for 11 normal posterior pdfs. These posterior pdfs use 2 levels of prior pdfs. The first prior pdf $h(\theta|\tau)$ is $N(0,\tau)$ where τ is a random variance hyperparameter. The distribution of τ is the inverse-gamma (IG) hyperprior pdf. $\tau \sim IG(2,1)$ where $IG(\alpha,\beta) \equiv \pi(\tau) = \frac{\beta^\alpha e^{-\beta/\tau}}{\Gamma(\alpha)\tau^{\alpha+1}}$. The likelihood function is $g(x|\theta) = N(\theta|\frac{1}{16})$. The 11 pdfs are posteriors for the observations $x = -4, -3.25, -1.75, -1, -0.25, 0.5, 1.25, 2.75$, and 3.5. The approximate posterior $F(\theta|x)$ uses a fuzzy approximation for the inverse-gamma hyperprior $\pi(\tau)$ (1000 uniform sample points on the support [0, 4], 15 rules, and 6000 learning iterations). The posterior pdfs show the distribution of θ given the data x.

We simulated a variation of the conjugate normal case. The likelihood is normally distributed with unknown mean $g(x|\theta) = N(\theta|\frac{1}{16})$. A normal prior pdf $h(\theta)$ models the unknown mean. We used a standard normal for $h(\theta)$ in the previous case. Here we assume $h(\theta)$ has unknown variance τ. So $h(\theta|\tau)$ is $N(0,\tau)$. We model τ with an inverse gamma (IG) hyperprior pdf: $\tau \sim IG(2,1)$ where $IG(\alpha,\beta) = \pi(\tau) = \frac{\beta^\alpha e^{-\beta/\tau}}{\Gamma(\alpha)\tau^{\alpha+1}}$. The inverse gamma prior is conjugate to the normal likelihood and so the resulting posterior is inverse gamma. Thus we have conjugacy in both the mean and variance parameters.

We obtain an approximation $F(\theta|x)$ for the posterior $f(\theta|x)$ by fuzzy approximation of the truncated hyperprior $\pi(\tau)$. Figure 9 shows how an adaptive sinc SAM approximates the truncated hyperprior. This fuzzy approximation used 1000 uniform sample points on the support [0, 4], 15 rules, and 6000 learning iterations.

Figure 10 shows the final fuzzy approximations for 11 normal posterior pdfs using this technique. The 11 pdfs are posteriors for the observations $x = -4$, -3.25, -2.5, -1.75, -1, -0.25, 0.5, 1.25, 2, 2.75, and 3.5. The posterior pdfs show the distribution of θ given the data x. We integrate τ out of $f(\theta|\tau|x)$ to yield the marginal posterior $f(\theta|x)$.

V. Doubly Fuzzy Bayesian Inference: Uniform Approximation

We will use the term *doubly fuzzy* to describe Bayesian inference where separate fuzzy systems $H(\theta)$ and $G(x|\theta)$ approximate the respective prior pdf $h(\theta)$ and the likelihood pdf $g(x|\theta)$. Theorem 3 below shows that the resulting fuzzy approximator F of the posterior pdf $f(\theta|x)$ still has the convex-sum structure (8) of a SAM fuzzy system.

The doubly fuzzy posterior approximator F requires only $m_1 m_2$ rules if the fuzzy likelihood approximator G uses m_1 rules and if the fuzzy prior approximator H uses m_2 rules. The $m_1 m_2$ if-part sets of F have a corresponding product structure as do the other fuzzy-system parameters. Corollary 3.1 shows that using an exact 2-rule representation reduces the corresponding rule number m_1 or m_2 to two. This is a tractable growth in rules for a single Bayesian inference. But the same structure leads in general to an exponential growth in posterior-approximator rules if the old posterior approximator becomes the new prior approximator in iterated Bayesian inference.

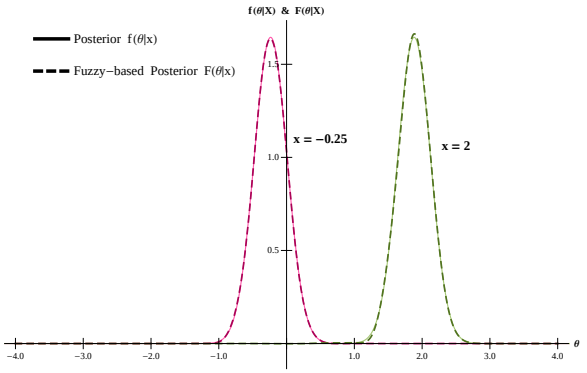

Fig. 11. Doubly fuzzy Bayesian inference: comparison of two normal posteriors and their doubly fuzzy approximators. The doubly fuzzy approximations use fuzzy prior-pdf approximator $H(\theta)$ and fuzzy likelihood-pdf approximator $G(x|\theta)$. The sinc-SAM fuzzy approximator $H(\theta)$ uses 15 rules to approximate the normal prior $h(\theta) = N(0,1)$. The Gaussian-SAM fuzzy likelihood approximator $G(x|\theta)$ uses 15 rules to approximate the two likelihood functions $g(x|\theta) = N(-0.25, \frac{1}{16})$ and $g(x|\theta) = N(2, \frac{1}{16})$. The two fuzzy approximators used 6000 learning iterations based on 500 uniform sample points.

Figure 11 shows the result of doubly fuzzy Bayesian inference for two normal posterior pdfs. A 15-rule Gaussian SAM G approximates two normal likelihood pdfs while a 15-rule sinc SAM H approximates a standard normal prior pdf.

We call the next theorem the Bayesian Approximation Theorem (BAT). The BAT shows that doubly fuzzy systems

can uniformly approximate posterior pdfs under some mild conditions. The proof derives an approximation error bound for $F(\theta|x)$ that does not depend on θ or x. Thus $F(\theta|x)$ uniformly approximates $f(\theta|x)$. The BAT holds in general for any uniform approximators of the prior or likelihood. Corollary 2.1 shows how the centroid and convex-sum structure of SAM fuzzy approximators H and G specifically bound the posterior approximator F. Theorem 3 gives further insight into the induced SAM structure of the doubly fuzzy posterior approximator F.

The statement and proof of the BAT require the following notation. Let $\mathcal{D}$ denote the set of all θ and let $\mathcal{X}$ denote the set of all x. Assume that $\mathcal{D}$ and $\mathcal{X}$ are compact. The prior is $h(\theta)$ and the likelihood is $g(x|\theta)$. $H(\theta)$ is a 1-dimensional SAM fuzzy system that uniformly approximates $h(\theta)$ in accord with the Fuzzy Approximation Theorem [13], [15]. $G(x|\theta)$ is a 2-dimensional SAM that uniformly approximates $g(x|\theta)$. Define the Bayes factors as $q(x) = \int_{\mathcal{D}} h(\theta)g(x|\theta)d\theta$ and $Q(x) = \int_{\mathcal{D}} H(\theta)G(x|\theta)d\theta$. Assume that $q(x) > 0$ so that the posterior $f(\theta|x)$ is well-defined for any sample data x. Let ΔZ denote the approximation error $Z - z$ for an approximator Z.

Theorem 2: *Bayesian Approximation Theorem.* Suppose that $h(\theta)$ and $g(x|\theta)$ are bounded and continuous and that $H(\theta)G(x|\theta) \neq 0$ almost everywhere. Then the doubly fuzzy SAM system $F(\theta|x) = HG/Q$ uniformly approximates $f(\theta|x)$ for all $\epsilon > 0 : \ |F(\theta|x) - f(\theta|x)| < \epsilon$.

The BAT proof in the Appendix also shows how sequences of uniform approximators H_n and G_n lead to a sequence of posterior approximators F_n that converges uniformly to F. Suppose we have such sequences H_n and G_n that uniformly approximate the respective prior h and likelihood g. Suppose $\epsilon_{h,n+1} < \epsilon_{h,n}$ and $\epsilon_{g,n+1} < \epsilon_{g,n}$ for all n. Define $F_n = \frac{H_n G_n}{\int H_n G_n}$. Then for all $\epsilon > 0$ there exists an $n_0 \in \mathbb{N}$ such that for all $n > n_0 : \quad |F_n(\theta|x) - F(\theta|x)| < \epsilon$ for all θ and for all x. The positive integer n_0 is the first n such that $\epsilon_{h,n}$ and $\epsilon_{g,n}$ satisfy (101). Hence F_n converges uniformly to F.

Corollary 2.1 below reveals the fuzzy structure of the BAT's uniform approximation when the prior H and likelihood G are uniform SAM approximators. The corollary shows how the convex-sum and centroidal structure of H and G produce centroid-based bounds on the fuzzy posterior approximator F. Recall first that Theorem 1 states that $F(\theta|x) = \sum_{j=1}^{m} p_j(\theta)c'_j(x|\theta)$ where $c'_j(x|\theta) = \frac{c_j g(x|\theta)}{\sum_{i=1}^{m} \int_{\mathcal{D}} g(x|u)p_i(u)c_i du}$. Replace the likelihood $g(x|\theta)$ with its doubly fuzzy SAM approximator $G(x|\theta)$ to obtain the posterior

$$F(\theta|x) \quad = \quad \sum_{j=1}^{m} p_j(\theta)C'_j(x|\theta) \tag{57}$$

where the then-part set centroids are

$$C'_j(x|\theta) \quad = \quad \frac{c_{h,j}G(x|\theta)}{\sum_{i=1}^{m} \int_{\mathcal{D}} G(x|u)p_i(u)c_{h,i}du}. \tag{58}$$

The $\{c_{h,k}\}_k$ are the then-part set centroids for the prior SAM approximator $H(\theta)$. $G(x|\theta)$ likewise has then-part set centroids $\{c_{g,j}\}_j$. Each SAM is a convex sum of its centroids from (48). This convex-sum structure induces bounds on H and G that in turn produce bounds on F. We next let the subscripts *max* and *min* denote the respective maximal and minimal centroids. The maximal centroids are positive. But the minimal centroids may be negative even though h and g are non-negative functions. We also assume that the minimal centroids are positive. So define the maximal and minimal product centroids as

$$c_{gh,max} \quad = \quad \max_{j,k} c_{g,j}c_{h,k} \quad = \quad c_{g,max}c_{h,max} \tag{59}$$

$$c_{gh,min} \quad = \quad \min_{j,k} c_{g,j}c_{h,k} \quad = \quad c_{g,min}c_{h,min}. \tag{60}$$

Then the BAT gives the following SAM-based bound.

Corollary 2.1: Centroid-based bounds for the doubly fuzzy posterior F.

Suppose that the set D of all θ has positive Lebesgue measure. Then the centroids of the H and G then-part sets bound the posterior F:

$$\frac{c_{gh,min}}{m(D)c_{gh,max}} \quad \leq \quad F(\theta|x) \quad \leq \quad \frac{c_{gh,max}}{m(D)c_{gh,min}}. \tag{61}$$

The size of the bounding interval depends on the size of the set D and on the minimal centroids of H and G. The lower bound is more sensitive to minimal centroids than the upper bound because dividing by a maximum is more stable than dividing by a minimum close to zero. The bounding interval becomes $[0, \infty)$ if any of the minimal centroids for H or G equal zero. The infinite bounding interval $[0, \infty)$ corresponds to the least informative case.

Similar centroid bounds hold for the multidimensional case. Suppose that the SAM-based posterior F is the multidimensional approximator $F : R \to R^p$ with $p > 1$. Then the same argument applies to the components of the centroids along each dimension. There are p bounding intervals $\frac{c^s_{gh,min}}{m(\mathcal{D})c^s_{gh,max}} \leq F_s(\theta|x) \leq \frac{c^s_{gh,max}}{m(\mathcal{D})c^s_{gh,min}}$ for each dimension s of the range R^p. These componentwise intervals define a bounding hypercube $\prod_{s=1}^{p}[\frac{c^s_{gh,min}}{m(\mathcal{D})c^s_{gh,max}}, \frac{c^s_{gh,max}}{m(\mathcal{D})c^s_{gh,min}}] \subset R^p$ for F.

The next theorem shows that a doubly fuzzy system's posterior $F(\theta|x)$ maintains the convex-sum structure (8) and has $m_1 m_2$ rules if the likelihood approximator G has m_1 rules and the prior approximator H has m_2 rules.

Theorem 3: Doubly fuzzy posterior approximators are SAMs with product rules.

Suppose an m_1-rule SAM fuzzy system $G(x|\theta)$ approximates (or represents) a likelihood pdf $g(x|\theta)$ and another m_2-rule SAM fuzzy system $H(\theta)$ approximates (or represents) a prior $h(\theta)$ pdf with m_2 rules:

$$G(x|\theta) \ = \ \frac{\sum_{j=1}^{m_1} w_{g,j}a_{g,j}(\theta)V_{g,j}c_{g,j}}{\sum_{i=1}^{m_1} w_{g,j}a_{g,j}(\theta)V_{g,j}} = \sum_{j=1}^{m_1} p_{g,j}(\theta)c_{g,j} \tag{62}$$

$$H(\theta) \ = \ \frac{\sum_{j=1}^{m_2} w_{h,j}a_{h,j}(\theta)V_{h,j}c_{h,j}}{\sum_{j=1}^{m_2} w_{h,j}a_{h,j}(\theta)V_{h,j}} = \sum_{j=1}^{m_2} p_{h,j}(\theta)c_{h,j} \tag{63}$$

where $p_{g,j}(\theta) = \frac{w_{g,j}a_{g,j}(\theta)V_{g,j}}{\sum_{i=1}^{m_1} w_{g,j}a_{g,j}(\theta)V_{g,j}}$ and $p_{h,j}(\theta) = \frac{w_{h,j}a_{h,j}(\theta)V_{h,j}}{\sum_{i=1}^{m_2} w_{h,j}a_{h,j}(\theta)V_{h,j}}$ are convex coefficients: $\sum_{j=1}^{m_1} p_{g,j}(\theta) = 1$ and $\sum_{j=1}^{m_2} p_{h,j}(\theta) = 1$. Then (a) and (b) hold:

(a) The fuzzy posterior approximator $F(\theta|x)$ is a SAM system with $m = m_1 m_2$ rules:

$$F(\theta|x) \quad = \quad \frac{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}\, c_{F,i}}{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}}. \tag{64}$$

(b) The m if-part set functions $a_{F,i}(\theta)$ of the fuzzy posterior approximator $F(\theta|x)$ are the products of the likelihood approximator's if-part sets $a_{g,j}(\theta)$ and the prior approximator's if-part sets $a_{h,k}(\theta)$:

$$a_{F,i}(\theta) \quad = \quad a_{g,j}(\theta)a_{h,k}(\theta). \tag{65}$$

for $i = m_2(j-1)+k$, $j = 1, \dots, m_1$, and $k = 1, \dots, m_2$. The weights w_{F_i}, then-part set volumes V_{F_i}, and centroids c_{F_i} also have the same likelihood-prior product form:

$$w_{F_i} \quad = \quad w_{g,j}w_{h,k} \tag{66}$$

$$V_{F_i} \quad = \quad V_{g,j}V_{h,k} \tag{67}$$

$$c_{F_i} \quad = \quad \frac{c_{g,j}c_{h,k}}{Q(x)}. \tag{68}$$

So the updated fuzzy system $F(\theta|x)$ has $m = m_1 m_2$ rules with weights $w_{F,i} = w_{g,j}w_{h,k}$, if-parts set functions $a_{F,i}(\theta) = a_{g,j}(\theta)a_{h,k}(\theta)$, then-part set volumes $V_{F,i} = V_{g,j}V_{h,k}$, and centroids $c_{F,i} = c_{g,j}c_{h,k}$ where $i = m_2(j-1)+k$, $j = 1, \dots, m_1$, and $k = 1, \dots, m_2$. Note that the m_1-rule fuzzy system $G(x|\theta)$ represents (or approximates) $g(x|\theta)$ as a function of θ when x is an observation.

VI. Conclusion

Fuzzy systems allow users to encode prior and likelihood information through fuzzy rules rather than through only a handful of closed-form probability densities. This can produce more accurate priors and likelihoods based on expert input or sample data or both. Gradient-descent learning algorithms specifically allow fuzzy systems to learn and tune rules based on the same type of collateral data that an expert might consult or that a statistical hypothesis might use. Different learning algorithms should produce different bounds on the fuzzy prior or likelihood approximations and those in turn should lead to different bounds on the fuzzy posterior approximation. Hierarchical Bayes systems can model hyperpriors with fuzzy approximators or with other "intelligent" learning systems such as neural networks or semantic networks. An open research problem is how to reduce the exponential rule explosion that doubly fuzzy Bayesian systems face in general in Bayesian iterative inference.

References

[1] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Volume I*, Prentice Hall, 2nd edition, 2001.

[2] P. Billingsley, *Probability and Measure*, John Wiley & Sons, 3rd edition, p. 269, 1995.

[3] B. P. Carlin and T. A. Louis, *Bayesian Methods for Data Analysis*, CRC Press, 3nd edition, 2009.

[4] M. H. DeGroot, *Optimal Statistical Decisions*, McGraw-Hill, 1970.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2nd edition, 2001.

[6] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Wiley-Interscience, 2nd edition, 1999.

[7] S. Frühwirth-Schnatter, "On fuzzy Bayesian inference," *Fuzzy Sets and Systems*, vol. 60, no. 1, pp. 41–58, 1993.

[8] R. V. Hogg, J. W. McKean, and A. T. Craig, *Introduction to Mathematical Statistics*, Prentice Hall, 6th edition, 2005.

[9] R. V. Hogg and E. A. Tanis, *Probability and Statistical Inference*, Prentice Hall, 7th edition, 2006.

[10] Y. Jin, "Fuzzy Modeling of High-dimensional Systems: Complexity Reduction and Interpretability Improvement," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 2, pp. 212–221, April 2000.

[11] B. Kosko, "Fuzzy Entropy and Conditioning," *Information Sciences*, vol. 40, pp. 165–174, 1986.

[12] B. Kosko, *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence*, Prentice Hall, 1992.

[13] B. Kosko, "Fuzzy Systems as Universal Approximators," *IEEE Transactions on Computers*, vol. 43, no. 11, pp. 1329–1333, November 1994.

[14] B. Kosko, "Optimal Fuzzy Rules Cover Extrema," *International Journal of Intelligent Systems*, vol. 10, no. 2, pp. 249–255, February 1995.

[15] B. Kosko, *Fuzzy Engineering*, Prentice Hall, 1996.

[16] B. Kosko, "Probable Equality, Superpower Sets, and Superconditionals," *International Journal of Intelligent Systems*, vol. 19, pp. 1151–1171, December 2004.

[17] B. Kosko and S. Mitaim, "Robust Stochastic Resonance: Signal Detection and Adaptation in Impulsive Noise," *Physical Review E*, vol. 64, no. 051110, October 2001.

[18] B. Kosko and S. Mitaim, "Stochastic Resonance in Noisy Threshold Neurons," *Neural Networks*, vol. 16, no. 5-6, pp. 755–761, June-July 2003.

[19] I. Lee, B. Kosko, and W. F. Anderson, "Modeling Gunshot Bruises in Soft Body Armor with Adaptive Fuzzy Systems," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 35, no. 6, pp. 1374–1390, December 2005.

[20] S. Mitaim and B. Kosko, "Neural Fuzzy Agents for Profile Learning and Adaptive Object Matching," *Presence*, vol. 7, no. 6, pp. 617–637, December 1998.

[21] S. Mitaim and B. Kosko, "The Shape of Fuzzy Sets in Adaptive Function Approximation," *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 4, pp. 637–656, August 2001.

[22] S. Mitra and S. K. Pal, "Fuzzy Self-Organization, Inferencing, and Rule Generation," *IEEE Transactions on Systems, Man, Cybernetics–A*, vol. 26, no. 5, pp. 608–620, 1996

[23] J. R. Munkres, *Topology*, Prentice Hall, 2nd edition, 2000.

[24] R. E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, 2004.

[25] C. L. Nikias and M. Shao, *Signal Processing with Alpha-Stable Distributions and Applications*, John Wiley & Sons, 1995.

[26] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*, John Wiley & Sons, 2000.

[27] S. Ross, *A First Course in Probability*, Prentice Hall, 7th edition, 2005.

[28] T. Terano, K. Asai, and M. Sugeno, *Fuzzy Systems Theory and Its Applications*, Academic Press, 1987.

[29] F. A. Watkins, *Fuzzy Engineering*, Ph.D. Dissertation, Department of Electrical Engineering, UC Irvine, 1994.

[30] F. A. Watkins, "The Representation Problem for Additive Fuzzy Systems," in *Proceedings of the IEEE International Conference on Fuzzy Systems (IEEE FUZZ-95)*, March 1995, vol. 1, pp. 117–122.

[31] R. Xu and D. C. Wunsch, *Clustering*, IEEE Press & Wiley, 2009.

[32] C. C. Yang, "Fuzzy Bayesian Inference," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 3, pp. 2707–2712, 1997.

[33] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

Appendix: Proofs of Theorems

This section restates the Theorems and provides proofs.

Theorem 1: The fuzzy posterior approximator is a SAM:

$$F(\theta|x) \quad = \quad \sum_{j=1}^{m} p_j(\theta)c'_j(x|\theta) \tag{69}$$

where the generalized then-part set centroids $c'_j(x|\theta)$ have the form

$$c'_j(\theta|x) = \frac{c_j(\theta)g(x|\theta)}{\sum_{i=1}^{m} \int_{\mathcal{D}} g(x|u)p_i(u)c_i(u)\,du} \tag{70}$$

for sample space $\mathcal{D}$.

Proof: The proof equates the fuzzy-based posterior $F(\theta|x)$ with the right-hand side of (2) and then expands according to Bayes Theorem:

$$F(\theta|x) = \frac{g(x|\theta)F(\theta)}{\int_{\mathcal{D}} g(x|u)F(u)du} \quad \text{by (2)} \quad (71)$$

$$= \frac{g(x|\theta)\sum_{j=1}^{m} p_j(\theta)c_j(\theta)}{\int_{\mathcal{D}} g(x|u)\sum_{j=1}^{m} p_j(u)c_j(u)\,du} \quad \text{by (8)} \quad (72)$$

$$= \frac{g(x|\theta)\sum_{j=1}^{m} p_j(\theta)c_j(\theta)}{\sum_{j=1}^{m}\int_{\mathcal{D}} g(x|u)p_j(u)c_j(u)\,du} \quad (73)$$

$$= \sum_{j=1}^{m} p_j(\theta)\left(\frac{c_j(\theta)g(x|\theta)}{\sum_{i=1}^{m}\int_{\mathcal{D}} g(x|u)p_i(u)c_i(u)\,du}\right) \quad (74)$$

$$= \sum_{j=1}^{m} p_j(\theta)c'_j(x|\theta). \quad \text{Q.E.D.} \quad (75)$$

Corollary 1.1: Suppose $g(x|\theta)$ approximates a Dirac delta function centered at x: $g(x|\theta) \approx \delta(\theta - x)$. Then $c'_j(\theta|x)$ in (70) becomes

$$c'_j(\theta|x) \approx \frac{c_j(\theta)g(x|\theta)}{F(x)}. \quad (76)$$

Proof: Suppose $g(x|\theta) \approx \delta(\theta - x)$. Then the integration in (70) becomes

$$\int_{\mathcal{D}} g(x|u)p_j(u)c_j(u)\,du \approx \int_{\mathcal{D}} \delta(u-x)p_j(u)c_j(u)\,du \quad (77)$$

$$= p_j(x)c_j(x). \quad (78)$$

Then (70) becomes

$$c'_j(\theta|x) \approx \frac{c_j(\theta)g(x|\theta)}{\sum_{i=1}^{m} p_i(x)c_i(x)} \quad (79)$$

$$= \frac{c_j(\theta)g(x|\theta)}{F(x)} \quad \text{by (8)} \quad \text{Q.E.D.} \quad (80)$$

Corollary 1.2: Suppose we can approximate the likelihood $g(x|\theta)$ with constant $g(x|m_j)$ and then-part set centroids $c_j(\theta)$ with constant $c_j(m_j)$ over $\mathcal{D}_{p_j}$. Then $c'_j(\theta|x)$ in (70) becomes

$$c'_j(\theta|x) \approx \frac{c_j(\theta)g(x|\theta)}{\sum_{i=1}^{m} g(x|m_i)U_{p_i}c_i(m_j)} \quad (81)$$

where $U_{p_j} = \int_{\mathcal{D}_{p_j}} p_j(u)du$.

Proof. Suppose $g(x|\theta) \approx g(x|m_j)$ and $c_j(\theta) \approx c_j(m_j)$ over $\mathcal{D}_{p_j}$. Then

$$\int_{\mathcal{D}} g(x|u)p_j(u)c_j(u)\,du \approx \int_{\mathcal{D}_{p_j}} g(x|m_j)p_j(u)c_j(m_j)\,du \quad (82)$$

$$= g(x|m_j)\,U_{p_j}c_j(m_j). \quad (83)$$

Then (70) becomes

$$c'_j(\theta|x) \approx \frac{c_j(\theta)g(x|\theta)}{\sum_{i=1}^{m} g(x|m_i)U_{p_i}c_i(m_j)} \quad \text{Q.E.D.} \quad (84)$$

Theorem 2: *Bayesian Approximation Theorem.* Suppose that $h(\theta)$ and $g(x|\theta)$ are bounded and continuous and that $H(\theta)G(x|\theta) \neq 0$ almost everywhere. Then the doubly fuzzy SAM system $F(\theta|x) = HG/Q$ uniformly approximates $f(\theta|x)$ for all $\epsilon > 0 : \ |F(\theta|x) - f(\theta|x)| < \epsilon$.

Proof: Write the posterior pdf $f(\theta|x)$ as $f(\theta|x) = \frac{h(\theta)g(x|\theta)}{q(x)}$ and its approximator $F(\theta|x)$ as $F(\theta|x) = \frac{H(\theta)G(x|\theta)}{Q(x)}$. The SAM approximations for the prior and likelihood functions are uniform [15]. So they have approximation error bounds ϵ_h and ϵ_g that do not depend on x or θ:

$$|\Delta H| < \epsilon_h \quad \text{and} \quad |\Delta G| < \epsilon_g \quad (85)$$

where $\Delta H = H(\theta) - h(\theta)$ and $\Delta G = G(x|\theta) - g(x|\theta)$. The posterior error ΔF is

$$\Delta F = F - f = \frac{HG}{Q(x)} - \frac{hg}{q(x)}. \quad (86)$$

Expand HG in terms of the approximation errors to get

$$HG = (\Delta H + h)(\Delta G + g) \quad (87)$$

$$= \Delta H\Delta G + \Delta Hg + h\Delta G + hg. \quad (88)$$

We have assumed that $HG \neq 0$ almost everywhere and so $Q \neq 0$. We now derive an upper bound for the Bayes-factor error $\Delta Q = Q - q$:

$$\Delta Q = \int_{\mathcal{D}} (\Delta H\Delta G + \Delta Hg + h\Delta G + hg - hg)\ d\theta. \quad (89)$$

So

$$|\Delta Q| \leq \int_{\mathcal{D}} |\Delta H\Delta G + \Delta Hg + h\Delta G|\ d\theta \quad (90)$$

$$\leq \int_{\mathcal{D}} \left(|\Delta H||\Delta G| + |\Delta H|g + h|\Delta G|\right)\ d\theta \quad (91)$$

$$< \int_{\mathcal{D}} (\epsilon_h\epsilon_g + \epsilon_h g + h\epsilon_g)\ d\theta \quad \text{by (85).} \quad (92)$$

Parameter set $\mathcal{D}$ has finite Lebesgue measure $m(\mathcal{D}) = \int_{\mathcal{D}} d\theta < \infty$ because $\mathcal{D}$ is a compact subset of a metric space and thus [23] it is (totally) bounded. Then the bound on ΔQ becomes

$$|\Delta Q| < m(\mathcal{D})\epsilon_h\epsilon_g + \epsilon_g + \epsilon_h \int_{\mathcal{D}} g(x|\theta)\ d\theta \quad (93)$$

because $\int_{\mathcal{D}} h(\theta)d\theta = 1$.

We now invoke the extreme value theorem [6]. The extreme value theorem states that a continuous function on a compact set attains both its maximum and minimum. The extreme value theorem allows us to use maxima and minima instead of suprema and infima. Now $\int_{\mathcal{D}} g(x|\theta)\ d\theta$ is a continuous function of x because $g(x|\theta)$ is a continuous nonnegative function. The range of $\int_{\mathcal{D}} g(x|\theta)\ d\theta$ is a subset of the right half line $(0, \infty)$ and its domain is the compact set $\mathcal{D}$. So $\int_{\mathcal{D}} g(x|\theta)\ d\theta$ attains a finite maximum value. Thus

$$|\Delta Q| < \epsilon_q \quad (94)$$

where we define the error bound ϵ_q as

$$\epsilon_q = m(\mathcal{D})\epsilon_h\epsilon_g + \epsilon_g + \epsilon_h \max_x \left\{\int_{\mathcal{D}} g(x|\theta)\ d\theta\right\}. \quad (95)$$

Rewrite the posterior approximation error ΔF as

$$\Delta F = \frac{qHG - Qhg}{qQ} \quad (96)$$

$$= \frac{q(\Delta H\Delta G + \Delta Hg + h\Delta G + hg) \ - \ Qhg}{q(q + \Delta Q)} \quad (97)$$

Inequality (94) implies that $-\epsilon_q < \Delta Q < \epsilon_q$ and that $(q - \epsilon_q) < (q + \Delta Q) < (q + \epsilon_q)$. Then (85) gives similar inequalities for ΔH and ΔG. So

$$\frac{q[-\epsilon_h\epsilon_g - \min(g)\epsilon_h - \min(h)\epsilon_g] \ - \ \epsilon_q hg}{q(q - \epsilon_q)} < \Delta F$$

$$< \frac{q[\epsilon_h\epsilon_g + \max(g)\epsilon_h + \max(h)\epsilon_g] \ + \ \epsilon_q hg}{q(q - \epsilon_q)}. \quad (98)$$

The extreme value theorem ensures that the maxima in (98) are finite. The bound on the approximation error ΔF does not depend on θ. But q still depends on the value of the data sample x. So (98) guarantees at best a pointwise approximation of $f(\theta|x)$ when x is arbitrary. We can improve the result by finding bounds for q that do not depend on x. Note that $q(x)$ is a continuous function of $x \in X$ because hg is continuous. So the extreme value theorem ensures that the Bayes factor q has a finite upper bound and a positive lower bound.

The term $q(x)$ attains its maximum and minimum by the extreme value theorem. The minimum of $q(x)$ is positive because we assumed $q(x) > 0$ for all x. Hölder's inequality gives $|q| \leq \left(\int_{\mathcal{D}} |h| d\theta\right) \left(\|g(x,\theta)\|_\infty\right) = \|g(x,\theta)\|_\infty$ since h is a pdf. So the maximum of $q(x)$ is finite because g is bounded: $0 < \min\{q(x)\} \leq \max\{q(x)\} < \infty$. Then

$$\epsilon_- < \Delta F < \epsilon_+ \tag{99}$$

if we define the error bounds ϵ_- and ϵ_+ as

$$\epsilon_- = \frac{(-\epsilon_h \epsilon_g - \min\{g\}\epsilon_h - \min\{h\}\epsilon_g)\min\{q\} - hg\epsilon_q}{\min\{q\}\left(\min\{q\} - \epsilon_q\right)} \tag{100}$$

$$\epsilon_+ = \frac{(\epsilon_h \epsilon_g + \max\{g\}\epsilon_h + \max\{h\}\epsilon_g)\max\{g\} + hg\epsilon_q}{\min\{q\}\left(\min\{q\} - \epsilon_q\right)} \tag{101}$$

Now $\epsilon_q \to 0$ as $\epsilon_g \to 0$ and $\epsilon_h \to 0$. So $\epsilon_- \to 0$ and $\epsilon_+ \to 0$. The denominator of the error bounds must be non-zero for this limiting argument. We can guarantee this when $\epsilon_q < \min\{q\}$. This condition is not restrictive because the functions h and g fix or determine q independent of the approximators H and G involved and because $\epsilon_q \to 0$ when $\epsilon_h \to 0$ and $\epsilon_g \to 0$. So we can achieve arbitrarily small ϵ_q that satisfies $\epsilon_q < \min\{q\}$ by choosing appropriate ϵ_h and ϵ_g. Then $\Delta F \to 0$ as $\epsilon_g \to 0$ and $\epsilon_h \to 0$. So $|\Delta F| \to 0$. Q.E.D.

Corollary 2.1: Centroid-based bounds for the doubly fuzzy posterior F.

Suppose that the set $\mathcal{D}$ of all θ has positive Lebesgue measure. Then the centroids of the H and G then-part sets bound the posterior F:

$$\frac{c_{gh,min}}{m(\mathcal{D})c_{gh,max}} \leq F(\theta|x) \leq \frac{c_{gh,max}}{m(\mathcal{D})c_{gh,min}}. \tag{102}$$

Proof: The convex-sum structure constrains the values of the SAMs: $H(\theta) \in [c_{h,min}, c_{h,max}]$ for all θ and $G(x|\theta) \in [c_{g,min}, c_{g,max}]$ for all x and θ. Then (58) implies

$$C'_j(x|\theta) \geq \frac{c_{gh,min}}{c_{gh,max}\sum_{i=1}^m \int_{\mathcal{D}} p_i(u)du} \tag{103}$$

$$= \frac{c_{gh,min}}{m(\mathcal{D})c_{gh,max}} \quad \text{for all } x \text{ and } \theta \tag{104}$$

since $\sum_{i=1}^m \int_{\mathcal{D}} p_i(u)du = \int_{\mathcal{D}} \sum_{i=1}^m p_i(u)du = \int_{\mathcal{D}} du = m(\mathcal{D})$ where $m(\mathcal{D})$ denotes the (positive) Lebesgue measure of $\mathcal{D}$. The same argument gives the upper bound:

$$C'_j(x|\theta) \leq \frac{c_{gh,max}}{m(\mathcal{D})c_{gh,min}} \tag{105}$$

for all x and θ. Thus (104) and (105) give bounds for all centroids:

$$\frac{c_{gh,min}}{m(\mathcal{D})c_{gh,max}} \leq C'_j(x|\theta) \leq \frac{c_{gh,max}}{m(\mathcal{D})c_{gh,min}} \tag{106}$$

for all x and θ. This bounding interval applies to $F(\theta|x)$ because the posterior approximator also has a convex-sum structure. Thus

$$\frac{c_{gh,min}}{m(\mathcal{D})c_{gh,max}} \leq F(\theta|x) \leq \frac{c_{gh,max}}{m(\mathcal{D})c_{gh,min}} \tag{107}$$

for all x and θ. Q.E.D.

Theorem 3: Doubly fuzzy posterior approximators are SAMs with product rules.

Suppose an m_1-rule SAM fuzzy system $G(x|\theta)$ approximates (or represents) a likelihood pdf $g(x|\theta)$ and another m_2-rule SAM fuzzy system $H(\theta)$ approximates (or represents) a prior $h(\theta)$ pdf with m_2 rules:

$$G(x|\theta) = \frac{\sum_{j=1}^{m_1} w_{g,j} a_{g,j}(\theta) V_{g,j} c_{g,j}}{\sum_{i=1}^{m_1} w_{g,j} a_{g,j}(\theta) V_{g,j}} = \sum_{j=1}^{m_1} p_{g,j}(\theta) c_{g,j} \tag{108}$$

$$H(\theta) = \frac{\sum_{j=1}^{m_2} w_{h,j} a_{h,j}(\theta) V_{h,j} c_{h,j}}{\sum_{j=1}^{m_2} w_{h,j} a_{h,j}(\theta) V_{h,j}} = \sum_{j=1}^{m_2} p_{h,j}(\theta) c_{h,j} \tag{109}$$

where $p_{g,j}(\theta) = \frac{w_{g,j} a_{g,j}(\theta) V_{g,j}}{\sum_{i=1}^{m_1} w_{g,j} a_{g,j}(\theta) V_{g,j}}$ and $p_{h,j}(\theta) = \frac{w_{h,j} a_{h,j}(\theta) V_{h,j}}{\sum_{i=1}^{m_2} w_{h,j} a_{h,j}(\theta) V_{h,j}}$ are convex coefficients: $\sum_{j=1}^{m_1} p_{g,j}(\theta) = 1$ and $\sum_{j=1}^{m_2} p_{h,j}(\theta) = 1$. Then (a) and (b) hold:

(a) The fuzzy posterior approximator $F(\theta|x)$ is a SAM system with $m = m_1 m_2$ rules:

$$F(\theta|x) = \frac{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}\, c_{F,i}}{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}}. \tag{110}$$

(b) The m if-part set functions $a_{F,i}(\theta)$ of the fuzzy posterior approximator $F(\theta|x)$ are the products of the likelihood approximator's if-part sets $a_{g,j}(\theta)$ and the prior approximator's if-part sets $a_{h,k}(\theta)$:

$$a_{F,i}(\theta) = a_{g,j}(\theta) a_{h,k}(\theta). \tag{111}$$

for $i = m_2(j-1) + k$, $j = 1, \ldots, m_1$, and $k = 1, \ldots, m_2$. The weights w_{F_i}, then-part set volumes V_{F_i}, and centroids c_{F_i} also have the same likelihood-prior product form:

$$w_{F_i} = w_{g,j} w_{h,k} \tag{112}$$

$$V_{F_i} = V_{g,j} V_{h,k} \tag{113}$$

$$c_{F_i} = \frac{c_{g,j} c_{h,k}}{Q(x)}. \tag{114}$$

Proof: The fuzzy system $F(\theta|x)$ has the form

$$F(\theta|x) = \frac{H(\theta)G(x|\theta)}{\int_{\mathcal{D}} H(t)G(x|t)\, dt} \tag{115}$$

$$= \frac{1}{Q(x)} \frac{\sum_{j=1}^{m_1} w_{g,j} a_{g,j}(\theta) V_{g,j} c_{g,j}}{\sum_{i=1}^{m_1} w_{g,j} a_{g,j}(\theta) V_{g,j}} \frac{\sum_{j=1}^{m_2} w_{h,j} a_{h,j}(\theta) V_{h,j} c_{h,j}}{\sum_{j=1}^{m_2} w_{h,j} a_{h,j}(\theta) V_{h,j}} \tag{116}$$

$$= \frac{\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} w_{g,j}\, w_{h,k}\, a_{g,j}(\theta) a_{h,k}(\theta)\, V_{g,j}\, V_{h,k}\, \frac{c_{g,j}\, c_{h,k}}{Q(x)}}{\sum_{j=1}^{m_1}\sum_{k=1}^{m_2} w_{g,j}\, w_{h,k}\, a_{g,j}(\theta) a_{h,k}(\theta)\, V_{g,j}\, V_{h,k}} \tag{117}$$

$$= \frac{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}\, c_{F,i}}{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}} \quad \text{Q.E.D.} \tag{118}$$