Fuzzy Optim Decis Making (2012) 11:241–268
DOI 10.1007/s10700-012-9130-0

Triply fuzzy function approximation for hierarchical Bayesian inference

Osonde Osoba · Sanya Mitaim · Bart Kosko

Received: 18 May 2011 / Accepted: 13 April 2012 / Published online: 1 May 2012

Abstract We prove that three independent fuzzy systems can uniformly approximate Bayesian posterior probability density functions by approximating the prior and likelihood probability densities as well as the hyperprior probability densities that underly the priors. This triply fuzzy function approximation extends the recent theorem for uniformly approximating the posterior density by approximating just the prior and likelihood densities. This approximation allows users to state priors and hyper-priors in words or rules as well as to adapt them from sample data. A fuzzy system with just two rules can exactly represent common closed-form probability densities so long as they are bounded. The function approximators can also be neural networks or any other type of uniform function approximator. Iterative fuzzy Bayesian inference can lead to rule explosion. We prove that conjugacy in the if-part set functions for prior, hyperprior, and likelihood fuzzy approximators reduces rule explosion. We also prove that a type of semi-conjugacy of if-part set functions for those fuzzy approximators results in fewer parameters in the fuzzy posterior approximator.

Keywords Adaptive fuzzy system · Bayesian inference · Function approximation · Hierarchical Bayes · Hyperprior

O. Osoba · B. Kosko (✉)
Department of Electrical Engineering, Signal and Image Processing Institute,
University of Southern California, Los Angeles, CA 90089-2564, USA
e-mail: kosko@usc.edu

O. Osoba
e-mail: osondeos@usc.edu

S. Mitaim
Department of Electrical and Computer Engineering, Faculty of Engineering,
Thammasat University, Pathumthani 12120, Thailand
e-mail: msanya@engr.tu.ac.th

1 Triply fuzzy function approximation for Bayesian posteriors

We extend a new theorem on uniform Bayesian approximation to the more general case that allows the use and approximation of hyperprior probability density functions (pdfs). We have recently shown (Osoba et al. 2011) that independent additive fuzzy systems H and G can uniformly approximate the respective Bayesian prior pdf $h(\theta)$ and the likelihood pdf $g(x|\theta)$ and thereby uniformly approximate the posterior pdf $f(\theta|x)$. The posterior pdf $f(\theta|x)$ arises from the usual Bayes Theorem (Kosko 2004) combination of the prior $h(\theta)$ and the likelihood $g(x|\theta)$:

$$f(\theta|x) = \frac{h(\theta)g(x|\theta)}{\int h(u)g(x|u)du}. \quad (1)$$

The last section shows that conjugate and what we call "semi-conjugate" fuzzy if-part sets can further simplify triply fuzzy function approximation.

The fuzzy posterior approximator F is an additive fuzzy system (Kosko 1992, 1996) that has the ratio form $F(\theta|x) = H(\theta)G(x|\theta)/Q(x)$. $H(\theta)$ is a 1-D standard additive model (SAM) fuzzy system that uniformly approximates $h(\theta)$ in accord with the Fuzzy Approximation Theorem (Kosko 1994, 1996). $G(x|\theta)$ is a 2-D SAM that uniformly approximates $g(x|\theta)$. The integral term $Q = \int_{\mathcal{D}} H(\theta)G(x|\theta)d\theta$ is the approximate Bayes factor that integrates over a compact set $\mathcal{D}$ of θ values. This Bayesian Approximation Theorem allows users to work with arbitrary priors and likelihoods based on if-then rules as well as based on familiar closed-form pdfs (Osoba et al. 2011). The Watkins Representation Theorem in Sect. 2 below further shows that an additive fuzzy system with just two rules can exactly represent any such closed-form pdf so long as the pdf is bounded. So fuzzy approximators substantially extend the practical and theoretical range of Bayesian statistical inference.

Figure 1 shows two simulation instances of this recent Bayesian Approximation Theorem (Osoba et al. 2011) and the resulting doubly fuzzy approximation of the posterior pdf. Each fuzzy SAM fuzzy system uses 15 rules. The prior approximator $H(\theta)$ uses sinc-shaped if-part fuzzy sets while the posterior approximator $G(x|\theta)$ uses Gaussian if-part fuzzy sets. The next section defines SAM fuzzy systems and shows how to adapt these two types of if-part fuzzy sets given training samples from the prior pdf $h(\theta)$ and the likelihood pdf $g(x|\theta)$. The samples can be noisy and the approximators can also use histograms to uniformly approximate the pdfs from noisy random draws from the pdfs (Osoba et al. 2011). The histograms uniformly approximate the pdfs in accord with the Glivenko–Cantelli Theorem (Billingsley 1995).

The simulations in Fig. 1 use a standard-normal prior $h(\theta) = N(0, 1)$ and the two different normal likelihoods $g(x|\theta) = N(-0.25, 1/16)$ and $g(x|\theta) = N(2, 1/16)$. The well-known conjugacy relation between normal priors and normal likelihoods yields a normal posterior pdf $f(\theta|x)$ (Bickel and Doksum 2001; Carlin and Louis 2009; DeGroot 1970). Iterative Bayesian inference takes the current posterior as the new prior in the next round of Bayesian inference from new likelihood data. Conjugacy relations greatly simplify this process but also unduly restrict the choice of Bayesian priors and likelihoods and thus of posteriors.

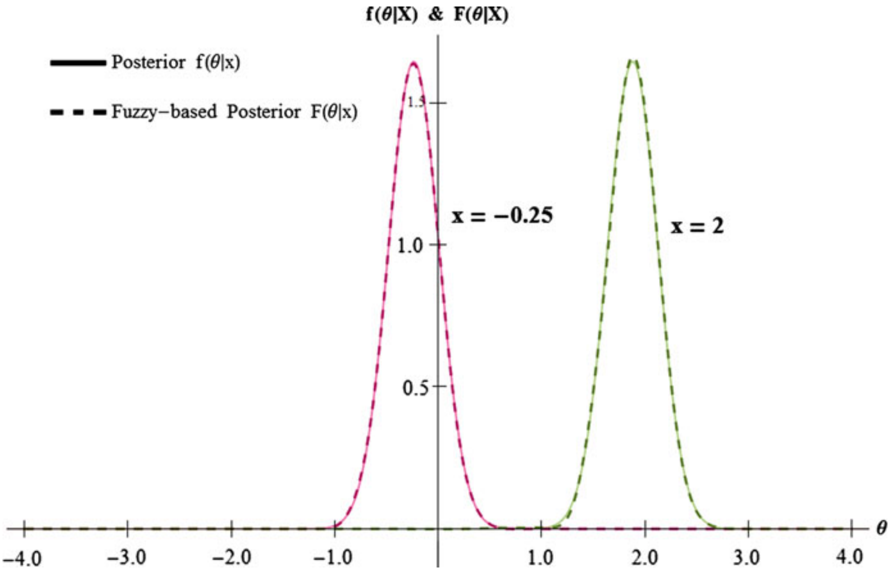

Fig. 1 Doubly fuzzy Bayesian inference: comparison of two normal posteriors and their doubly fuzzy approximators. The fuzzy approximators use Gaussian and sinc set functions. The doubly fuzzy approximations use fuzzy prior-pdf approximator $H(\theta)$ and fuzzy likelihood-pdf approximator $G(x|\theta)$. 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 $g(x|\theta) = N(\theta, \frac{1}{16})$ for $x = -0.25$ and $x = 2$. The two fuzzy approximators used 6,000 learning iterations based on 500 uniform sample points

We next state the Bayesian Approximation Theorem (Osoba et al. 2011) for sake of completeness and for comparison with Theorem 2 below.

Theorem 1 *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 $\epsilon > 0$: $|F(\theta|x) - f(\theta|x)| < \epsilon$ for all x and all θ.*

Our goal is to extend the Bayesian Approximation Theorem to allow the prior pdf h to depend on its own uncertain parameter τ through hyperprior pdf $\pi(\tau) \sim h(\theta|\tau)$ where random variable τ has pdf $\pi(\tau)$. This hierarchical Bayes case implies a more complex posterior with an extra parameter dimension:

$$f(\theta, \tau|x) \sim g(x|\theta)\, h(\theta|\tau)\, \pi(\tau). \tag{2}$$

Integrating over τ removes the extra parameter dimension and gives back the original posterior pdf:

$$f(\theta|x) \sim \int g(x|\theta)\, h(\theta|\tau)\, \pi(\tau)\, d\tau. \tag{3}$$

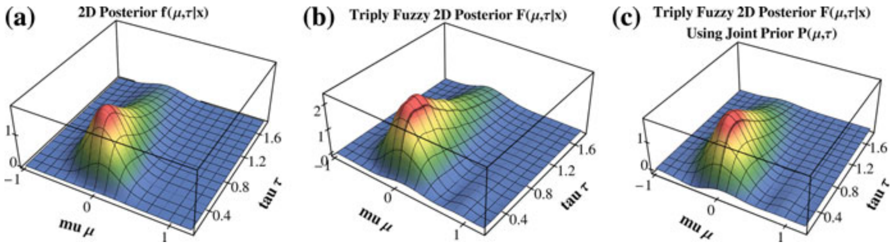

Fig. 2 Triply fuzzy Bayesian inference: comparison of a 2-D posterior $f(\mu, \tau|x) \propto g(x|\mu)h(\mu|\tau)\pi(\tau)$ and its triply fuzzy approximator $F(\mu, \tau|x)$. The *first panel* (**a**) shows the approximand $f(\mu, \tau|x)$. The *second panel* (**b**) shows a triply fuzzy approximator $F(\mu, \tau|x)$ that used a 2-D fuzzy approximation $H(\mu|\tau)$ for the conditional prior $h(\mu|\tau)$ and a 1-D fuzzy approximation $\Pi(\tau)$ for the hyperprior pdf $\pi(\tau)$ and a 1-D fuzzy likelihood-pdf approximator $G(x|\mu)$. The *third panel* (**c**) shows a triply fuzzy approximator $F(\mu, \tau|x)$ that used a 2-D fuzzy approximation $P(\mu, \tau) = (H \times \Pi)(\mu, \tau)$ for the joint prior $p(\mu, \tau) = (h \times \pi)(\mu, \tau)$. The likelihood approximation is the same as in the *second panel*. The sinc-SAM fuzzy approximators $H(\mu|\tau)$ and $P(\mu, \tau)$ use 6 rules to approximate the respective 2-D pdfs $h(\mu|\tau) = N(1, \tau)$ and $h(\mu|\tau)\pi(\tau) = N(1, \tau)IG(2, 1)$. The hyperprior Gaussian-SAM approximator $\Pi(\tau)$ used 12 rules to approximate an inverse-gamma pdf $\pi(\tau) = IG(2, 1)$. The Gaussian-SAM fuzzy likelihood approximator $G(x|\mu)$ used 15 rules to approximate the likelihood function $g(x|\mu) = N(\mu, \frac{1}{16})$ for $x = -0.25$. The 2-D conditional prior fuzzy approximator $H(\mu|\tau)$ used 15,000 learning iterations based on 6,000 uniform sample points. The hyperprior fuzzy approximator $\Pi(\tau)$ used 6,000 iterations on 120 uniform sample points. The likelihood fuzzy approximator used 6,000 iterations based on 500 uniform sample points

Theorem 2 below achieves this extension by adding a third SAM system $\Pi(\tau)$ to approximate the hyperprior $\pi(\tau)$. This involves triply fuzzy function approximation of the posterior. So we call the theorem the Extended Bayesian Approximation Theorem. The proof of this theorem in the Appendix is quite general and does not depend on the structure of the uniform fuzzy approximators. So the approximators can be neural networks or polynomials or any other uniform function approximators.

Figure 2 shows a simulation instance of triply fuzzy function approximation in accord with the Extended Bayesian Approximation Theorem. It shows that the 2-D fuzzy approximator $F(\mu, \tau|x)$ approximates the posterior pdf $f(\mu, \tau|x) \propto g(x|\mu)h(\mu|\tau)\pi(\tau)$ for hierarchical Bayesian inference. The sample data x is normal. A normal prior distribution $h(\mu|\tau) = N(1, \tau)$ models the population mean μ of the data. An inverse gamma $IG(2, 1)$ hyperprior models the variance τ of the prior. An inverse gamma hyperprior $\pi(\tau) = IG(\alpha, \beta)$ has the form $\pi(\tau) = e^{-\frac{\beta}{\tau}} \left(\frac{\beta}{\tau}\right)^{\alpha} \Big/ \tau\Gamma(\alpha)$ for $\tau > 0$ where Γ is the gamma function. The posterior fuzzy approximator $F(\mu, \tau|x)$ is proportional to the triple-product approximator $G(x|\mu)H(\mu|\tau)\Pi(\tau)$. These three adaptive SAMs separately approximate the three corresponding Bayesian pdfs: $G(x|\mu)$ approximates the 1-D likelihood pdf $g(x|\mu)$. $H(\mu|\tau)$ approximates the 2-D conditional prior pdf $h(\mu|\tau)$. And $\Pi(\tau)$ approximates the 1-D hyperprior pdf $\pi(\tau)$.

Figure 2 also shows a simulation instance where the posterior approximator $F(\mu, \tau|x)$ uses a single 2-D approximator $P(\mu, \tau)$ for the joint prior pdf $p(\mu, \tau) = h(\mu|\tau)\pi(\tau)$ instead of a separate 2-D approximator $H(\mu|\tau)$ for $h(\mu|\tau)$ and a separate 1-D approximator $\Pi(\tau)$ for $\pi(\tau)$. Both fuzzy posterior approximators $F(\mu, \tau|x) \propto G(x|\mu)H(\mu|\tau)\Pi(\tau)$ and $F(\mu, \tau|x) \propto G(x|\mu)P(\mu, \tau)$ quickly and uniformly approximate the posterior pdf $f(\mu, \tau|\theta)$.

2 Adaptive additive fuzzy systems

This section reviews the SAM fuzzy systems (Kosko 1994, 1995, 1996; Mitaim and Kosko 2001). A key property of a SAM fuzzy system F is that it represents the output $F(\theta)$ as a convex combination of the centroids of the then-part fuzzy sets in the system's if-then rules.

2.1 SAM fuzzy systems

A SAM fuzzy system computes the output $F(\theta)$ by taking the centroid of the sum of the "fired" or scaled then-part sets: $F(\theta) = Centroid(w_1 a_1(\theta) B_1 + \cdots + w_m a_m(\theta) B_m)$. Then the SAM Theorem states that the output $F(\theta)$ is a simple convex-weighted sum of the then-part set centroids c_j (Kosko 1992, 1994, 1996; Mitaim and Kosko 2001):

$$F(\theta) = \frac{\sum_{j=1}^{m} w_j a_j(\theta) V_j c_j}{\sum_{j=1}^{m} w_j a_j(\theta) V_j} = \sum_{j=1}^{m} p_j(\theta) c_j. \tag{4}$$

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 $p_1(\theta), \ldots, p_m(\theta)$ 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 $[-0.217, 1]$. They require some care in simulations because the denominator in (4) 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 (4) 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 in Mitaim and Kosko (2001): sinc, Gaussian, triangle, Cauchy, Laplace, and generalized hyperbolic tangent. The simulations below use sinc and Gaussian if-part sets. 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. Sinc if-part sets often produce the approximators that converges fastest in simulations (Mitaim and Kosko 2001).

SAMs define "model-free" statistical estimators in the following sense (Kosko 1996; Lee et al. 2005; Mitaim and Kosko 2001):

$$E[Y|\Theta = \theta] = F(\theta) = \sum_{j=1}^{m} p_j(\theta) c_j \tag{5}$$

$$V[Y|\Theta = \theta] = \sum_{j=1}^{m} p_j(\theta)\sigma^2_{B_j} + \sum_{j=1}^{m} p_j(\theta)[c_j - F(\theta)]^2. \tag{6}$$

The additive fuzzy structure yields the conditional expectation in (5). The SAM additive structure (4) yields the more specific form of the conditional variance in (6). Neither conditional moment involves any assumptions of joint probability structure such as joint Gaussianity. Nor do they reflect a particular state model such as a linear model. They are in this sense model-free estimators.

The first term on the right of (6) measures the inherent uncertainty in the m then-part rules. The second term is an interpolation penalty. It uses the normalized rule-firing weight $p_j(\theta)$ to weight how much the fuzzy system's output $F(\theta)$ resembles the centroid c_j of the jth then-part set B_j. Relations (5) and (6) generalize the usual unconditional mean and variance of mixture densities (Hogg et al. 2005) both because of their conditional structure and because (5) and (6) expressly depend on the current input θ.

The fuzzy applications in Lee et al. (2005) plot both the conditional expectation or $F(\theta)$ surfaces and the corresponding conditional variance surfaces. This paper does not plot the conditional variances because it focuses just on first-order function approximation. We thus ignore the attendant second-order uncertainty of the rules used in the function approximation.

A SAM fuzzy system F can always approximate a function 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 (Jin 2000; Kosko 1995; Mitra and Pal 1996). Optimal rules tend to reside at the extrema or turning points of the approximand f and so optimal fuzzy rules "patch the bumps" (Kosko 1995). 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 (Mitaim and Kosko 1998, 2001). The respective prior, hyperprior, and likelihood approximators $H : R^2 \to R$, $G : R^2 \to R$, and $\Pi : R \to R$ require at most $\mathcal{O}(k^2)$ rules and thus do not suffer rule explosion. But Theorem 3 below shows that iterative Bayesian inference can produce its own rule explosion (Osoba et al. 2011).

2.2 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 $F = f$ (Watkins 1994, 1995). 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 $V_1 = V_2$:

$$F(\theta) = \frac{\sum_{j=1}^{2} w_j a_j(\theta) V_j c_j}{\sum_{j=1}^{2} w_j a_j(\theta) V_j} \tag{7}$$

$$= \frac{a(\theta)c_1 + a^c(\theta)c_2}{a(\theta) + a^c(\theta)} \tag{8}$$

$$= f(\theta) \tag{9}$$

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

The representation technique builds f directly into the structure of the two if-then rules. A constant f needs a SAM with only one rule. Let $h(\theta)$ be any bounded prior pdf such as the beta $\beta(8, 5)$ pdf. 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

$$a(\theta) = \frac{\sup h - h(\theta)}{\sup h - \inf h} = 1 - \frac{11^{11}}{7^7 4^4}\theta^7(1-\theta)^4 \tag{10}$$

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 as in Corollary 3.2.

2.3 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 $h(\theta_1), h(\theta_2), \ldots$ 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 or other tests for a candidate pdf. An ASAM can learn the prior pdf even from noisy random samples drawn from the pdf (Osoba et al. 2011). Unsupervised clustering techniques can also train an ASAM if there is sufficient cluster data (Kosko 1992, 1996; Xu 2009). The ASAM prior simulations in the next section show how H approximates $h(\theta)$ when the ASAM trains on 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 any SAM parameter m_j and error E in the generic gradient-descent algorithm (Kosko 1996; Mitaim and Kosko 2001)

$$m_j(t+1) = m_j(t) - \mu_t \frac{\partial E}{\partial m_j} \tag{11}$$

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

$$E(\theta) = \frac{1}{2}(f(\theta) - F(\theta))^2 = \frac{1}{2}\varepsilon(\theta)^2 \tag{12}$$

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:

$$\frac{\partial E}{\partial m_j} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial a_j}\frac{\partial a_j}{\partial m_j} \tag{13}$$

$$\frac{\partial E}{\partial c_j} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial c_j} \tag{14}$$

$$\frac{\partial E}{\partial V_j} = \frac{\partial E}{\partial F}\frac{\partial F}{\partial V_j} \tag{15}$$

with partial derivatives (Kosko 1996; Mitaim and Kosko 2001)

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

$$\frac{\partial F}{\partial a_j} = [c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}. \tag{17}$$

The SAM ratio (4) with equal rule weights $w_1 = \cdots = w_m$ gives (Kosko 1996; Mitaim and Kosko 2001)

$$\frac{\partial F}{\partial c_j} = \frac{a_j(\theta)V_j}{\sum_{i=1}^m a_i(\theta)V_i} = p_j(\theta) \tag{18}$$

$$\frac{\partial F}{\partial V_j} = \frac{a_j(\theta)[c_j - F(\theta)]}{\sum_{i=1}^m a_i(\theta)V_i} = [c_j - F(\theta)]\frac{p_j(\theta)}{V_j}. \tag{19}$$

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

$$c_j(t+1) = c_j(t) + \mu_t \varepsilon(\theta) p_j(\theta) \tag{20}$$

$$V_j(t+1) = V_j(t) + \mu_t \varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{V_j}. \tag{21}$$

The learning laws for the if-part set parameters follow in like manner by expanding $\frac{\partial a_j}{\partial m_j}$ in (13).

The simulations in Figs. 1, 2 tune the location m_j and the width or dispersion d_j parameters of the if-part set functions a_j for sinc and Gaussian if-part sets with the following learning laws.

2.3.1 Sinc ASAM learning law

Define the sinc function as

$$\text{sinc}(x) = \frac{\sin(x)}{x} \tag{22}$$

The sinc set function a_j has the form

$$a_j(\theta) = \text{sinc}\left(\frac{\theta - m_j}{d_j}\right) \tag{23}$$

with parameter learning laws (Kosko 1996; Mitaim and Kosko 2001)

$$m_j(t+1) = m_j(t) + \mu_t \varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}\left(a_j(\theta) - \cos\left(\frac{\theta - m_j}{d_j}\right)\right)\frac{1}{\theta - m_j} \tag{24}$$

$$d_j(t+1) = d_j(t) + \mu_t \varepsilon(\theta)[c_j - F(\theta)]\frac{p_j(\theta)}{a_j(\theta)}\left(a_j(\theta) - \cos\left(\frac{\theta - m_j}{d_j}\right)\right)\frac{1}{d_j}. \tag{25}$$

2.3.2 Sinc 2D ASAM learning law

The sinc set function a_j has the form

$$a_j(x, y) = \text{sinc}\left(\frac{x - m_{x,j}}{d_{x,j}}\right)\text{sinc}\left(\frac{y - m_{y,j}}{d_{y,j}}\right) \tag{26}$$

with parameter learning laws (Kosko 1996; Mitaim and Kosko 2001)

$$\begin{aligned} m_{x,j}(t+1) &= m_{x,j}(t) + \mu_t \varepsilon(x, y)[c_j - F(x, y)] \\ &\times \left(a_j(x, y) - \cos\left(\frac{x - m_{x,j}}{d_{x,j}}\right)\text{sinc}\left(\frac{y - m_{y,j}}{d_{y,j}}\right)\right)\frac{p_j(x, y)}{a_j(x, y)}\left(\frac{1}{x - m_{x,j}}\right) \end{aligned} \tag{27}$$

$$\begin{aligned} d_{x,j}(t+1) &= d_{x,j}(t) + \mu_t \varepsilon(x, y)[c_j - F(x, y)] \\ &\times \left(a_j(x, y) - \cos\left(\frac{x - m_{x,j}}{d_{x,j}}\right)\text{sinc}(\frac{y - m_{y,j}}{d_{y,j}})\right)\frac{p_j(x, y)}{a_j(x, y)}\left(\frac{1}{d_{x,j}}\right). \end{aligned} \tag{28}$$

2.3.3 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

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

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

The Gaussian learning laws have the same functional form in the 2-D case. We replace $a_j(\theta)$ with $a_j(x, y)$:

$$a_j(x, y) = \exp\left[-\left(\frac{x - m_{x,j}}{d_{x,j}}\right)^2 - \left(\frac{y - m_{y,j}}{d_{y,j}}\right)^2\right] \quad (32)$$

with parameter learning laws

$$m_{x,j}(t+1) = m_{x,j}(t) + \mu_t \varepsilon(x, y) p_j(x, y)[c_j - F(x, y)] \frac{x - m_{x,j}}{d_{x,j}^2} \quad (33)$$

$$d_{x,j}(t+1) = d_{x,j}(t) + \mu_t \varepsilon(x, y) p_j(x, y)[c_j - F(x, y)] \frac{(x - m_{x,j})^2}{d_{x,j}^3}. \quad (34)$$

3 Triply fuzzy approximation with hyperpriors

We use the term *triply fuzzy* to describe Bayesian inference where $\Pi(\tau)$, $H(\theta|\tau)$, and $G(x|\theta)$ are the respective uniform approximators for the hyperprior pdf $\pi(\tau)$, the prior pdf $h(\theta|\tau)$, and the likelihood pdf $g(x|\theta)$. The 2-D pdf $p(\theta, \tau) = h(\theta|\tau)\pi(\tau)$ describes the dependence between θ and τ.

The statement and proof of both Bayesian approximation theorems require the following notation.The hyperprior pdf is $\pi(\tau)$. The prior is $h(\theta|\tau)$ and the likelihood is $g(x|\theta)$. $P(\theta, \tau)$ is a 2-D SAM fuzzy system that uniformly approximates $p(\theta, \tau) = h(\theta|\tau)\pi(\tau)$ in accord with the Fuzzy Approximation Theorem (Kosko 1994). $G(x|\theta)$ is a 2-D SAM that uniformly approximates $g(x|\theta)$. Let $\mathcal{D}$ denote the set of all (θ, τ) and let $\mathcal{X}$ denote the set of all x. Assume that $\mathcal{D}$ and $\mathcal{X}$ are compact. Define the Bayes factors as $q(x) = \int_{\mathcal{D}} p(\theta, \tau) g(x|\theta) d\tau d\theta$ and $Q(x) = \int_{\mathcal{D}} P(\theta, \tau) G(x|\theta) d\tau d\theta$. Assume that $q(x) > 0$ so that the posterior $f(\theta, \tau|x)$ is well-defined for any sample data x.

We can now state the Extended Bayesian Approximation Theorem. The proof is in the "Appendix" and relies on the Extreme Value Theorem (Munkres 2000). It does not require that the uniform approximator be a fuzzy system. The vector structure of the proof also allows the hyperprior prior to depend on its own hyperprior and so on. Figure 2a shows the approximand or the original posterior pdf. Figure 2b shows the adapted triply fuzzy approximator of the posterior pdf. Figure 2c shows an adapted non-separable fuzzy approximator of the posterior.

Theorem 2 *Extended Bayesian Approximation Theorem. Suppose that* $h(\theta|\tau)$, $\pi(\tau)$, *and* $g(x|\theta)$ *are bounded and continuous. Suppose that* $\Pi(\tau)H(\theta|\tau)G(x|\theta) = P(\theta, \tau)G(x|\theta) \neq 0$ *almost everywhere. Then the triply fuzzy SAM system* $F(\theta, \tau|x) = PG/Q$ *uniformly approximates* $f(\theta, \tau|x)$ *for all* $\epsilon > 0$: $|F(\theta, \tau|x) - f(\theta, \tau|x)| < \epsilon$ *for all* x *and all* (θ, τ).

The proof of Theorem 2 also implies that an n-D fuzzy posterior approximator F uniformly approximates the posterior f.

Figure 3 shows another simulation instance of triply fuzzy function approximation. But this instance works with a non-conjugate arbitrary Bayesian model. It uses

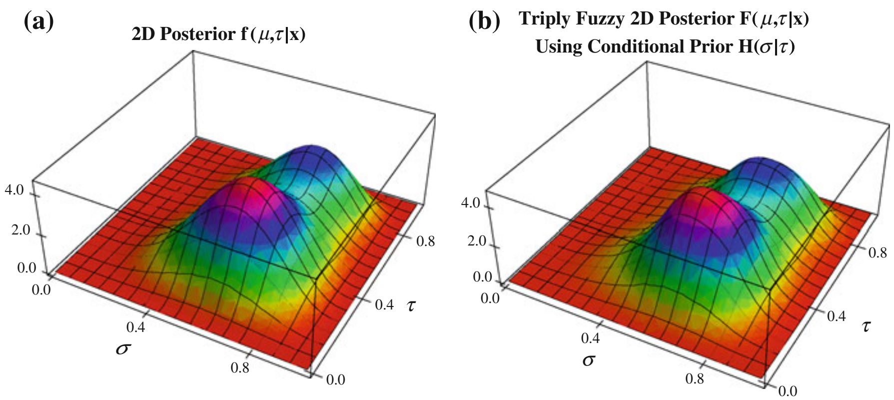

Fig. 3 Triply fuzzy Bayesian inference: comparison of a 2-D non-conjugate posterior $f(\sigma, \tau|x) \propto g(x|\sigma)h(\sigma|\tau)\pi(\tau)$ and its triply fuzzy approximator $F(\sigma, \tau|x)$. The *first panel* shows the approximand $f(\sigma, \tau|x)$. The *second panel* shows a triply fuzzy approximator $F(\sigma, \tau|x)$ that used a 2-D fuzzy approximation $H(\sigma|\tau)$ for the conditional prior $h(\sigma|\tau)$ and a 1-D fuzzy approximation $\Pi(\tau)$ for the hyperprior pdf $\pi(\tau)$ and a 1-D fuzzy likelihood-pdf approximator $G(x|\sigma)$. The Gaussian-SAM fuzzy approximator $H(\sigma|\tau)$ used 6 rules to approximate the 2-D pdf $h(\sigma|\tau) = \beta(6+2\tau, 4)$. The hyperprior Gaussian-SAM approximator $\Pi(\tau)$ used 12 rules to approximate a beta pdf $\pi(\tau) = \frac{1}{3}\beta(12, 4) + \frac{2}{3}\beta(4, 7)$. The Gaussian-SAM fuzzy likelihood approximator $G(x|\sigma)$ used 12 rules to approximate the likelihood function $g(x|\sigma) = N(0, \sigma)$ for $x = 0.25$. The 2-D conditional prior fuzzy approximator $H(\sigma|\tau)$ used 6,000 learning iterations based on 3,970 uniform sample points. The hyperprior fuzzy approximator $\Pi(\tau)$ used 15,000 iterations on 1,000 uniform sample points. The likelihood fuzzy approximator $G(x|\sigma)$ used 15,000 iterations based on 300 uniform sample points

normal data with unknown standard deviation σ. A conditional prior $h(\sigma|\tau) = \beta(6 + 2\tau, 4)$ models the distribution of the unknown standard deviation. A hyperprior $\pi(\tau) = \frac{1}{3}\beta(12, 4) + \frac{2}{3}\beta(4, 7)$ models the τ parameter of the conditional prior $h(\sigma|\tau)$. The hyperprior $\pi(\tau) = \frac{1}{3}\beta(12, 4) + \frac{2}{3}\beta(4, 7)$ is a bimodal mixture of two β pdfs. The 2-D fuzzy approximator $F(\sigma, \tau|x)$ approximates the posterior pdf $f(\sigma, \tau|x) \propto g(x|\sigma)h(\sigma|\tau)\pi(\tau)$ for this arbitrary model. The posterior fuzzy approximator $F(\sigma, \tau|x)$ is again proportional to the triple-product approximator $G(x|\sigma)H(\sigma|\tau)\Pi(\tau)$. $G(x|\sigma)$ approximates the 1-D likelihood pdf $g(x|\sigma)$. $H(\sigma|\tau)$ approximates the 2-D conditional prior pdf $h(\sigma|\tau)$. And $\Pi(\tau)$ approximates the 1-D hyperprior pdf $\pi(\tau)$.

4 Semi-conjugacy of the if-part sets in the fuzzy posterior approximator

4.1 SAM posterior and if-part conjugacy

This section explores conjugacy and semi-conjugacy effects on the if-part setfunctions involved in doubly and triply fuzzy Bayesian inferences. We restate Theorem 3 for the doubly fuzzy case in Osoba et al. (2011) that shows that updates preserve the SAM structure but with exponentially increasing rules and extend it to triply fuzzy and any n-many fuzzy cases. Theorem 4 and Corollaries 4.1–4.4 show that updates also preserve the shapes of the if-part sets (semi-conjugacy of if-part set functions)

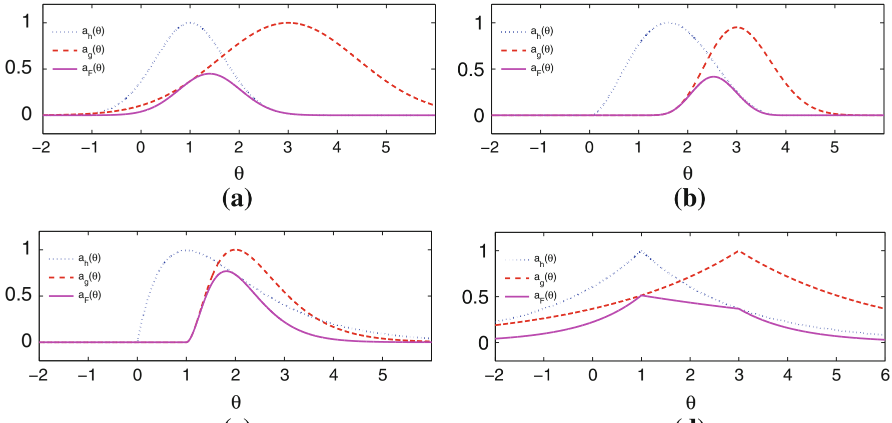

Fig. 4 Conjugacy and semi-conjugacy of the doubly fuzzy posterior if-part set functions $a_F(\theta) = a_h(\theta)a_g(x|\theta)$. **a** Gaussian if-part set functions have the form of (77) where $a_h(\theta) = G(1, 1, 1; \theta)$ and $a_g(\theta) = G(3, 2, 1; \theta)$ give Gaussian $a_F(\theta) = G(\frac{7}{5}, \frac{4}{5}, e^{-4/5}; \theta)$. **b** beta if-part set functions have the form of (93) where $a_h(\theta) = B(0, 4, 2, 3, 29; \theta)$ and $a_g(\theta) = B(1, 6, 6, 12, 9 \times 10^4; \theta)$ give semi-beta $a_F(\theta)$. **c** gamma if-part set functions have the form of (102) where $a_h = G(0, 1, 2, 3, 2.7; \theta)$ and $a_g = G(1, 1, 2, 0.5, 7.4; \theta)$ give semi-gamma $a_F(\theta)$. **d** Laplace if-part set functions have the form of (108) where $a_h(\theta) = L(1, 2; \theta)$ and $a_g(\theta) = L(3, 3; \theta)$ give semi-Laplace $a_F(\theta)$

if both SAM fuzzy systems in the doubly fuzzy cases use if-part set functions that belong to conjugate families in Bayesian statistics. The result also holds for triply fuzzy systems. Figure 4 shows examples of such if-part sets in Corollaries 4.1–4.4. The conjugacy of Gaussian if-part sets is straightforward. The conjugacy of the beta, gamma, and Laplace if-part sets is only partial (semi-conjugacy) because we cannot combine the functions' exponents and because two beta set functions or two gamma set functions need not share the same supports.

Theorem 3 *Preservation of SAM structure in fuzzy Bayesian inference:*

(i) 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:

$$G(x|\theta) = \frac{\sum_{j=1}^{m_1} w_{g,j}a_{g,j}(\theta)V_{g,j}c_{g,j}}{\sum_{i=1}^{m_1} w_{g,j}a_{g,j}(\theta)V_{g,j}} = \sum_{j=1}^{m_1} p_{g,j}(\theta)c_{g,j} \tag{35}$$

$$H(\theta) = \frac{\sum_{j=1}^{m_2} w_{h,j}a_{h,j}(\theta)V_{h,j}c_{h,j}}{\sum_{j=1}^{m_2} w_{h,j}a_{h,j}(\theta)V_{h,j}} = \sum_{j=1}^{m_2} p_{h,j}(\theta)c_{h,j} \tag{36}$$

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:

$$F(\theta|x) = \frac{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}\, c_{F,i}}{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}}. \tag{37}$$

(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 $a_{h,j}(\theta)$:

$$a_{F,i}(\theta) = a_{g,j}(\theta) a_{h,k}(\theta). \tag{38}$$

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:

$$w_{F_i} = w_{g,j} w_{h,k} \tag{39}$$

$$V_{F_i} = V_{g,j} V_{h,k} \tag{40}$$

$$c_{F_i} = \frac{c_{g,j} c_{h,k}}{Q(x)}. \tag{41}$$

(ii) Triply fuzzy posterior approximators and n-many 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 $g(x|\theta)$, an m_2-rule SAM fuzzy system $H(\theta,\tau)$ approximates (or represents) a prior pdf $h(\theta|\tau)$ with m_2 rules, an m_3-rule SAM fuzzy system $\Pi(\theta)$ approximates (or represents) a hyper-prior pdf $pi(\tau)$ with m_3 rules:

$$G(x|\theta) = \frac{\sum_{j=1}^{m_1} w_{g,j} a_{g,j}(\theta) V_{g,j} c_{g,j}}{\sum_{i=1}^{m_1} w_{g,j} a_{g,j}(\theta) V_{g,j}} = \sum_{j=1}^{m_1} p_{g,j}(\theta) c_{g,j} \tag{42}$$

$$H(\theta,\tau) = \frac{\sum_{j=1}^{m_2} w_{h,j} a_{h,j}(\theta,\tau) V_{h,j} c_{h,j}}{\sum_{j=1}^{m_2} w_{h,j} a_{h,j}(\theta,\tau) V_{h,j}} = \sum_{j=1}^{m_2} p_{h,j}(\theta,\tau) c_{h,j} \tag{43}$$

$$\Pi(\tau) = \frac{\sum_{j=1}^{m_3} w_{\pi,j} a_{\pi,j}(\tau) V_{h,j} c_{h,j}}{\sum_{j=1}^{m_3} w_{\pi,j} a_{\pi,j}(\tau) V_{\pi,j}} = \sum_{j=1}^{m_3} p_{h,j}(\tau) c_{h,j} \tag{44}$$

where $p_{g,j}(\theta) = \frac{w_{g,j} a_{g,j}(\theta) V_{g,j}}{\sum_{i=1}^{m_1} w_{g,i} a_{g,i}(\theta) V_{g,i}}$, $p_{h,j}(\theta,\tau) = \frac{w_{h,j} a_{h,j}(\theta,\tau) V_{h,j}}{\sum_{i=1}^{m_2} w_{h,i} a_{h,i}(\theta,\tau) V_{h,i}}$, and $p_{\pi,j}(\tau) = \frac{w_{\pi,j} a_{\pi,j}(\tau) V_{\pi,j}}{\sum_{i=1}^{m_3} w_{\pi,i} a_{pi,i}(\tau) V_{h,i}}$ are convex coefficients: $\sum_{j=1}^{m_1} p_{g,j}(\theta) = 1$, $\sum_{j=1}^{m_2} p_{h,j}(\theta,\tau) = 1$, and $\sum_{j=1}^{m_2} p_{\pi,j}(\theta) = 1$. Then (a) and (b) hold:

(a) The fuzzy posterior approximator $F(\theta,\tau|x)$ is a SAM system with $m = m_1 m_2 m_3$ rules:

$$F(\theta,\tau|x) = \frac{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}\, c_{F,i}}{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}}. \tag{45}$$

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

$$a_{F,i}(\theta, \tau) = a_{g,j}(\theta) a_{h,k}(\theta, \tau) a_{\pi,l}(\tau) \tag{46}$$

for $i = l + m_3(k-1) + m_2 m_3(j-1)$, $j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$, and $l = 1, \ldots, m_3$. The weights w_{F_i}, then-part set volumes V_{F_i}, and centroids c_{F_i} also have the same likelihood-prior-hyper-prior product form:

$$w_{F_i} = w_{g,j} w_{h,k} w_{\pi,l} \tag{47}$$

$$V_{F_i} = V_{g,j} V_{h,k} V_{\pi,l} \tag{48}$$

$$c_{F_i} = \frac{c_{g,j} c_{h,k} c_{\pi,l}}{Q(x)} \tag{49}$$

where $Q(x) = \int_{\mathcal{D}} G(x|\theta) H(\theta, \tau) \Pi(\tau)\, d\tau d\theta$.

This implies that the n-many fuzzy posterior approximators are also SAMs with product rules.

Proof Doubly fuzzy case.

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

$$F(\theta|x) = \frac{H(\theta) G(x|\theta)}{\int_{\mathcal{D}} H(t) G(x|t)\, dt} \tag{50}$$

$$= \frac{1}{Q(x)} \left(\sum_{j=1}^{m_1} p_{g,j}(\theta)\, c_{g,j} \right) \left(\sum_{j=1}^{m_2} p_{h,j}(\theta)\, c_{h,j} \right) \tag{51}$$

$$= \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} p_{g,j}(\theta) p_{h,k}(\theta)\, \frac{c_{g,j}\, c_{h,k}}{Q(x)} \tag{52}$$

$$= \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \frac{w_{g,j} a_{g,j}(\theta) V_{g,j}}{\sum_{i=1}^{m_1} w_{g,i} a_{g,i}(\theta) V_{g,i}} \frac{w_{h,k} a_{h,k}(\theta) V_{h,k}}{\sum_{i=1}^{m_2} w_{h,i} a_{h,i}(\theta) V_{h,i}} \frac{c_{g,j}\, c_{h,k}}{Q(x)} \tag{53}$$

$$= \frac{\sum_{j=1}^{m_1} \sum_{k=1}^{m_2} w_{g,j}\, w_{h,k}\, a_{g,j}(\theta) a_{h,k}(\theta)\, V_{g,j}\, V_{h,k}\, \frac{c_{g,j}\, c_{h,k}}{Q(x)}}{\sum_{j=1}^{m_1} \sum_{k=1}^{m_2} w_{g,j}\, w_{h,k}\, a_{g,j}(\theta) a_{h,k}(\theta)\, V_{g,j}\, V_{h,k}} \tag{54}$$

$$= \frac{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}\, c_{F,i}}{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}} \tag{55}$$

$$= \sum_{i=1}^{m} p_{F,i}(\theta)\, c_{F,i}. \tag{56}$$

□

Proof Triply fuzzy case.

The fuzzy system $F(\theta, \tau|x)$ has the form

$$F(\theta, \tau|x) = \frac{G(x|\theta)H(\theta, \tau)\Pi(\tau)}{\int_{\mathcal{D}_\theta \times \mathcal{D}_\tau} G(x|t)H(t, s)\Pi(s)\, dt\, ds} \tag{57}$$

$$= \frac{1}{Q(x)} \left(\sum_{j=1}^{m_1} p_{g,j}(\theta)\, c_{g,j} \right) \left(\sum_{j=1}^{m_2} p_{h,j}(\theta, \tau)\, c_{h,j} \right) \left(\sum_{j=1}^{m_3} p_{\pi,j}(\theta)\, c_{\pi,j} \right) \tag{58}$$

$$= \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} p_{g,j}(\theta) p_{h,k}(\theta, \tau) p_{\pi,l}(\tau)\, \frac{c_{g,j}\, c_{h,k} c_{\pi,l}}{Q(x)} \tag{59}$$

$$= \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \sum_{l=1}^{m_3} \frac{w_{g,j} a_{g,j}(\theta) V_{g,j}}{\sum_{i=1}^{m_1} w_{g,i} a_{g,i}(\theta) V_{g,i}}\, \frac{w_{h,k} a_{h,k}(\theta, \tau) V_{h,k}}{\sum_{i=1}^{m_2} w_{h,i} a_{h,i}(\theta, \tau) V_{h,i}} \tag{60}$$

$$\times \frac{w_{\pi,l} a_{\pi,l}(\tau) V_{\pi,k}}{\sum_{i=1}^{m_3} w_{\pi,i} a_{\pi,i}(\tau) V_{\pi,i}}\, \frac{c_{g,j}\, c_{h,k}\, c_{\pi,l}}{Q(x)}$$

$$= \frac{\sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \sum_{l=1}^{m_3}\, w_{g,j} w_{h,k} w_{\pi,k}\, a_{g,j}(\theta) a_{h,k}(\theta, \tau) a_{\pi,l}(\tau) V_{g,j} V_{h,k} V_{\pi,l}\, \frac{c_{g,j}\, c_{h,k}\, c_{\pi,l}}{Q(x)}}{\sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \sum_{l=1}^{m_3}\, w_{g,j} w_{h,k} w_{\pi,k}\, a_{g,j}(\theta) a_{h,k}(\theta, \tau) a_{\pi,l}(\tau) V_{g,j} V_{h,k} V_{\pi,l}} \tag{61}$$

$$= \frac{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}\, c_{F,i}}{\sum_{i=1}^{m} w_{F,i}\, a_{F,i}(\theta)\, V_{F,i}} \tag{62}$$

□

Corollary 3.1 *Two-rule representation of* $g(x|\theta)$.

Suppose a 2-rule fuzzy system $G(x|\theta)$ represents a likelihood pdf $g(x|\theta)$ and an m-rule system $H(\theta)$ approximates the prior pdf $h(\theta)$. Then the fuzzy-based posterior (or "updated" system) $F(\theta|x)$ is a SAM fuzzy system with $2m$ rules.

Proof Suppose a 2-rule fuzzy system $G(x|\theta)$ represents a likelihood pdf $g(x|\theta)$:

$$G(x|\theta) = \sum_{j=1}^{2} p_{g,j}(\theta) c_{g,j} \; = \; \sum_{k=1}^{2} a_{g,j}(\theta) c_{g,j} \tag{63}$$

where the if-part set functions have the form (from the Watkins Representation Theorem)

$$a_{g,1}(x|\theta) = \frac{g(x|\theta) - \inf g(x|\theta)}{\sup g(x|\theta) - \inf g(x|\theta)} \tag{64}$$

$$a_{g,2}(x|\theta) = a^c_{g,1}(\theta) \;=\; 1 - a_{g,1}(x|\theta) \tag{65}$$

$$= \frac{\sup g(x|\theta) - g(x|\theta)}{\sup g(x|\theta) - \inf g(x|\theta)} \tag{66}$$

and the centroids are $c_{g,1} = \sup g$ and $c_{g,2} = \inf g$. And suppose that an m-rule fuzzy system $H(\theta)$ with equal weights $w_i = \cdots = w_m$ and volumes $V_i = \cdots = V_m$ approximates (or represents) the prior $h(\theta)$. Then (37) becomes

$$F(\theta|x) = \frac{\sum_{j=1}^{m}\sum_{k=1}^{2} a_{g,k}(x|\theta)\, a_{h,j}(\theta)\, \frac{c_{g,k}\, c_{h,j}}{Q(x)}}{\sum_{j=1}^{m}\sum_{k=1}^{2} a_{g,k}(x|\theta)\, a_{h,j}(\theta)} \tag{67}$$

$$= \frac{\sum_{j=1}^{m} a_{g,1}(x|\theta)a_{h,j}(\theta)\, \frac{c_{g,1}c_{h,j}}{Q(x)} + a_{g,2}(x|\theta)a_{h,j}(\theta)\, \frac{c_{g,2}c_{h,j}}{Q(x)}}{\sum_{j=1}^{m} a_{g,1}(x|\theta)\, a_{h,j}(\theta) + a_{g,2}(x|\theta)\, a_{h,j}(\theta)} \tag{68}$$

$$= \frac{\sum_{j=1}^{m} a_{g,1}(x|\theta)a_{h,j}(\theta)\, \frac{c_{g,1}c_{h,j}}{Q(x)} + (1 - a_{g,1}(x|\theta))a_{h,j}(\theta)\, \frac{c_{g,2}c_{h,j}}{Q(x)}}{\sum_{j=1}^{m} a_{g,1}(x|\theta)\, a_{h,j}(\theta) + (1 - a_{g,1}(x|\theta))\, a_{h,j}(\theta)} \tag{69}$$

□

The above results imply that the number m of rules of a fuzzy system $F(\theta|x)$ after n stages will be $m_1 m_2^n = 2^n m$ rules. So the iterative fuzzy posterior approximator will in general suffer from exponential rule explosion.

At least one practical special case avoids this exponential rule explosion and produces only a linear or quadratic growth in fuzzy-posterior rules in iterative Bayesian inference. Suppose that we can keep track of past data involved in the Bayesian inference and that $g(x_1, \ldots, x_n|\theta) = g(\bar{x}_n|\theta)$. Then we can compute the likelihood pdf $g(\bar{x}_{n-1}|\theta)$ from $g(\bar{x}_n|\theta)$ for any new data x_n. Then we can update the original prior $H(\theta)$ and keep the number of rules at $2m$ (or m^2) if the fuzzy system uses two rules (or m rules).

Corollary 3.2 *Two-rule representation of both $h(\theta)$ and $g(x|\theta)$.*

Suppose a 2-rule fuzzy system $G(x|\theta)$ represents a likelihood function $g(x|\theta)$ and a 2-rule system $H(\theta)$ represents the prior $h(\theta)$. Then the fuzzy-based posterior $F(\theta|x)$ is a SAM fuzzy system with 4 (2×2) rules.

Proof Suppose a 2-rule fuzzy system $G(x|\theta)$ represents a likelihood pdf $g(x|\theta)$ as in (63–66). The 2-rule fuzzy system $H(\theta)$ likewise represents the prior pdf $h(\theta)$:

$$H(\theta) = \sum_{k=1}^{2} p_{h,k}(\theta)c_{h,k} \;=\; \sum_{k=1}^{2} a_{h,k}(\theta)c_{h,k}. \tag{70}$$

The Watkins Representation Theorem implies that the if-part set functions have the form

$$a_{h,1}(\theta) = \frac{h(\theta) - \inf h(\theta)}{\sup h(\theta) - \inf h(\theta)} \tag{71}$$

$$a_{h,2}(\theta) = a^c_{h,1}(\theta) \;=\; 1 - a_{h,1}(\theta) \tag{72}$$

$$= \frac{\sup h(\theta) - h(\theta)}{\sup h(\theta) - \inf h(\theta)} \tag{73}$$

with centroids $c_{h,1} = \sup h$ and $c_{h,2} = \inf h$. Then the SAM posterior $F(\theta|x)$ in (37) represents $f(\theta|x)$ with 4 rules:

$$F(\theta|x) = \frac{\sum_{j=1}^{2}\sum_{k=1}^{2} a_{g,j}(x|\theta)\, a_{h,k}(\theta)\, \frac{c_{g,j}\, c_{h,k}}{Q(x)}}{\sum_{j=1}^{2}\sum_{k=1}^{2} a_{g,j}(x|\theta)\, a_{h,k}(\theta)} \tag{74}$$

$$= \sum_{j=1}^{2}\sum_{k=1}^{2} a_{g,j}(x|\theta)\, a_{h,k}(\theta)\, \frac{c_{g,j}\, c_{h,k}}{q(x)} \tag{75}$$

$$= \sum_{i=1}^{4} a_{F,i}(\theta)\, c_{F,i} \tag{76}$$

because $\sum a_{g,j}(x|\theta) = \sum a_{h,k}(\theta) = 1$ and $Q(x) = q(x)$ in (74). □

Figure 5 shows the if-part sets $a_{h,k}(\theta)$ of the 2-rule SAM $H(\theta)$ that represents the beta prior $h(\theta) \sim \beta(9, 9)$ and the if-part sets $a_{g,j}(\theta)$ of the 2-rule SAM $G(x|\theta)$ that represents the binomial likelihood $g(20|\theta) \sim bin(20, 80)$. The resulting SAM posterior $F(\theta|20)$ that represents $f(\theta|20) \sim \beta(29, 69)$ has four rules with if-part sets $a_{F,i}(\theta) = a_{g,j}(\theta)a_{h,k}(\theta)$. The next theorem gives the main result on the conjugacy structure of doubly and triply fuzzy systems.

Theorem 4 *Conjugate fuzzy set functions.*

(i) The if-part sets of a doubly fuzzy posterior approximator are conjugate to the if-part sets of the fuzzy prior approximator. The product fuzzy if-part set functions $a_{F,i}(\theta)$ in Theorem 3.i(b) have the same functional form as the if-part prior set functions $a_{h,k}$ if $a_{h,k}$ is conjugate to the if-part likelihood set function $a_{g,j}$.
(ii) The if-part sets of a triply fuzzy posterior approximator are conjugate to the if-part sets of the fuzzy prior approximator. The product fuzzy if-part set functions $a_{F,i}(\theta)$ in Theorem 3.ii(b) have the same functional form as the if-part prior set functions $a_{h,k}$ if $a_{h,k}$ is conjugate to the if-part likelihood set function $a_{g,j}$ and if-part likelihood set function $a_{\pi,l}$.

Proof The product $a_{F,i}(\theta) = a_{g,j}(\theta)a_{h,k}(\theta)$ of two conjugate functions $a_{g,j}$ and $a_{h,k}$ will still have the same functional form as $a_{g,j}(\theta)$ and $a_{h,k}(\theta)$. Then the n parameters $\alpha_1, \ldots, \alpha_n$ define the if-part likelihood set function: $a_{g,j}(\theta) = f(\alpha_1, \ldots, \alpha_n; \theta)$. The n parameters $\beta_1, \ldots, \beta_n$ likewise define the if-part prior set function $a_{h,k}(\theta)$ with the

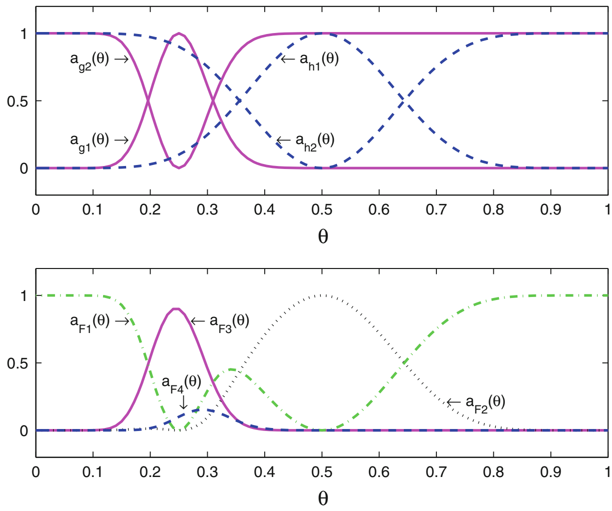

Fig. 5 Doubly fuzzy posterior representation. *Top*: two if-part sets $a_{g,j}(\theta)$ of the two-rule SAM likelihood representation $G(x|\theta) = g(20|\theta) \sim bin(20, 80)$ and two if-part sets $a_{h,k}(\theta)$ of the 2-rule SAM prior representation $H(x|\theta) = h(\theta) \sim \beta(9, 9)$. *Bottom*: four if-part sets $a_{F,i}(\theta) = a_{g,j}(\theta)a_{h,k}(\theta)$ of the 4-rule SAM posterior representation $F(\theta|x) = f(\theta|x)$

same functional form: $a_{h,k}(\theta) = f(\beta_1, \ldots, \beta_n; \theta)$. Then $a_{F,i}(\theta)$ also has the same functional form f given the n parameters $\gamma_1, \ldots, \gamma_n$: $a_{F,i}(\theta) = f(\gamma_1, \ldots, \gamma_n; \theta)$ where $\gamma_l = g_l(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_n)$ for $l = 1 \ldots, n$ for some functions $g_1, \ldots, g_n$ that do not depend on θ. □

Gaussian if-part sets are self-conjugate because of their exponential structure.

Corollary 4.1 *Conjugacy of Gaussian if-part sets.*

(i) Doubly fuzzy case.

Suppose that the SAM-based prior $H(\theta)$ uses Gaussian if-part sets $a_{h,k}(\theta) = G(m_{h,k}, d_{h,k}, \nu_{h,k}; \theta)$ and the SAM-based likelihood $G(x|\theta)$ also uses Gaussian if-part sets $a_{g,j}(\theta) = G(m_{g,j}, d_{g,j}, \nu_{g,j}; \theta)$ where

$$G(m, d, \nu; \theta) = \nu e^{-(\theta - m)^2/d^2} \tag{77}$$

for some positive constant $\nu > 0$. Then $F(\theta|x)$ in (37) will have set functions $a_{F,i}(\theta)$ that are also Gaussian:

$$a_{F,i}(\theta) = a_{g,j}(\theta)\, a_{h,k}(\theta) \tag{78}$$
$$= \nu_{F,i} e^{-(\theta - m_{F,i})^2/d_{F,i}^2} \tag{79}$$
$$= G(m_{F,i}, d_{F,i}, \nu_{F,i}; \theta) \tag{80}$$

where

$$m_{F,i} = \frac{d_{g,j}^2 m_{h,k} + d_{h,k}^2 m_{g,j}}{d_{g,j}^2 + d_{h,k}^2} \tag{81}$$

$$d_{F,i}^2 = \frac{d_{g,j}^2 d_{h,k}^2}{d_{g,j}^2 + d_{h,k}^2} \tag{82}$$

$$\nu_{F,i} = \nu_{h,k}\nu_{g,j} \exp\{-\frac{(m_{h,k} - m_{g,j})^2}{d_{g,j}^2 + d_{h,k}^2}\}. \tag{83}$$

for $j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$, and $i = m_2(j-1) + k$.

(ii) Triply fuzzy case.

Suppose that the SAM-based prior $H(\theta, \tau)$ uses factorable (product) Gaussian if-part sets $a_{h\theta k}(\theta, \tau) = G(m_{h\theta k}, d_{h\theta k}, \nu_{h\theta k}; \theta) G(m_{h\tau k}, d_{h\tau k}, \nu_{h\tau k}; \tau)$, the SAM-based likelihood $G(x|\theta)$ uses Gaussian if-part sets $a_{g\theta j}(\theta) = G(m_{g\theta j}, d_{g\theta j}, \nu_{g\theta j}; \theta)$, and the SAM-based hyper-prior $\Pi(\tau)$ also uses Gaussian if-part sets $a_{h\tau l}(\tau) = G(m_{h\tau l}, d_{h\tau l}, \nu_{h\tau l}; \tau)$. Then $F(\theta, \tau|x)$ in (45) will have set functions $a_{F,i}(\theta, \tau)$ that are products of two Gaussian sets:

$$a_{F,i}(\theta, \tau) = a_{g,j}(\theta) a_{h,k}(\theta, \tau) a_{\pi,l}(\tau) \tag{84}$$

$$= \nu_{F\theta i} e^{-(\theta - m_{F\theta i})^2 / d_{F\theta i}^2} \nu_{F\tau i} e^{-(\tau - m_{F\tau i})^2 / d_{F\tau i}^2} \tag{85}$$

$$= G(m_{F\theta i}, d_{F\theta i}, \nu_{F\theta i}; \theta) G(m_{F\tau i}, d_{F\tau i}, \nu_{F\tau i}; \tau) \tag{86}$$

where

$$m_{F\theta i} = \frac{d_{g\theta j}^2 m_{h\theta k} + d_{h\theta k}^2 m_{g\theta j}}{d_{g\theta j}^2 + d_{h\theta k}^2} \tag{87}$$

$$d_{F\theta i}^2 = \frac{d_{g\theta j}^2 d_{h\theta k}^2}{d_{g\theta j}^2 + d_{h\theta k}^2} \tag{88}$$

$$\nu_{F\theta i} = \nu_{h\theta k} \nu_{g\theta j} \exp\{-\frac{(m_{h\theta k} - m_{g\theta j})^2}{d_{g\theta j}^2 + d_{h\theta k}^2}\} \tag{89}$$

$$m_{F\tau i} = \frac{d_{\pi\tau l}^2 m_{h\tau k} + d_{h\tau k}^2 m_{\pi\tau l}}{d_{\pi\tau l}^2 + d_{h\tau k}^2} \tag{90}$$

$$d_{F\tau i}^2 = \frac{d_{\pi\tau l}^2 d_{h\tau k}^2}{d_{\pi\tau l}^2 + d_{h\tau k}^2} \tag{91}$$

$$\nu_{F\tau i} = \nu_{h\tau k} \nu_{\pi\tau l} \exp\{-\frac{(m_{h\tau k} - m_{\pi\tau l})^2}{d_{\pi\tau l}^2 + d_{h\tau k}^2}\}. \tag{92}$$

for $j = 1, \ldots, m_1$, $k = 1, \ldots, m_2$, $l = 1, \ldots, m_3$, and $i = l + m_3(k-1) + m_2 m_3 (j-1)$.

Corollary 4.1 also shows that if the fuzzy approximator $H(\theta, \tau)$ uses product if part set functions $a_h(\theta, \tau) = a_{h\theta}(\theta)a_{h\tau}(\tau)$ then the fuzzy posterior $F(\theta, \tau|x)$ also has product if-part sets $a_F(\theta, \tau) = a_{F\theta}(\theta)a_{F\tau}(\tau)$. This holds for higher dimension fuzzy approximators for Bayesian inference. Thus the corollaries below only state the results for doubly fuzzy cases.

Corollary 4.2 *Semi-conjugacy of beta if-part sets.*

Suppose that the SAM-based prior $H(\theta)$ uses beta (or binomial) if-part sets $a_{h,k}(\theta) = B(m_{h,k}, d_{h,k}, \alpha_{h,k}, \beta_{h,k}, \nu_{h,k}; \theta)$ and the SAM-based likelihood $G(x|\theta)$ also uses beta (or binomial) if-part sets $a_{g,j}(\theta) = B(m_{g,j}, d_{g,j}, \alpha_{g,j}, \beta_{g,j}, \nu_{g,j}; \theta)$ where

$$B(m, d, \alpha, \beta, \nu; \theta) = \nu\Big(\frac{\theta - m}{d}\Big)^{\alpha}\Big(1 - (\frac{\theta - m}{d})\Big)^{\beta} \tag{93}$$

if $0 < \frac{\theta - m}{d} < 1$ and for some constant $\nu > 0$. Then the posterior $F(\theta|x)$ in (37) will have if-part set functions $a_{F,i}(\theta)$ of semi-beta form

$$a_{F,i}(\theta) = a_{g,j}(\theta)\, a_{h,k}(\theta) \tag{94}$$
$$= \nu_{F,i}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{h,k}}\Big(1 - (\frac{\theta - m_{h,k}}{d_{h,k}})\Big)^{\beta_{h,k}}\Big(\frac{\theta - m_{g,j}}{d_{g,j}}\Big)^{\alpha_{g,j}}\Big(1 - (\frac{\theta - m_{g,j}}{d_{g,j}})\Big)^{\beta_{g,j}} \tag{95}$$
$$= \nu_{F,i}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{h,k}+\alpha_{g,j}\lambda_{jk}(\theta)}\Big(1 - (\frac{\theta - m_{h,k}}{d_{h,k}})\Big)^{\beta_{h,k}+\beta_{g,j}\gamma_{jk}(\theta)} \tag{96}$$

if $0 < \frac{\theta - m_{h,k}}{d_{h,k}} < 1$ and $0 < \frac{\theta - m_{g,j}}{d_{g,j}} < 1$ or if $\theta \in (m_{h,k}, m_{h,k} + d_{h,k}) \cap (m_{g,j}, m_{g,j} + d_{g,j})$ where

$$\lambda_{jk}(\theta) = \log_{(\frac{\theta - m_{h,k}}{d_{h,k}})}\Big(\frac{\theta - m_{g,j}}{d_{g,j}}\Big) \tag{97}$$
$$\gamma_{jk}(\theta) = \log_{(1-\frac{\theta - m_{h,k}}{d_{h,k}})}\Big(1 - \frac{\theta - m_{g,j}}{d_{g,j}}\Big). \tag{98}$$

A special case occurs if $m_{h,k} = m_{g,j}$ and $d_{h,k} = d_{g,j}$. Then $a_{F,i}$ has the beta conjugate form:

$$a_{F,i}(\theta) = \nu_{F,i}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{F,i}}\Big(1 - \Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)\Big)^{\beta_{F,i}} \tag{99}$$
$$= B(m_{h,k}, d_{h,k}, \alpha_{F,i}, \beta_{F,i}, \nu_{F,i}; \theta) \tag{100}$$

if $0 < \frac{\theta - m_{h,k}}{d_{h,k}} < 1$. Here $\alpha_{F,i} = \alpha_{h,k} + \alpha_{g,j}$, $\beta_{F,i} = \beta_{h,k} + \beta_{g,j}$, and $\nu_{F,i} = \nu_{h,k}\nu_{g,j}$.

The if-part fuzzy sets of the posterior approximation in (96) have beta-like form but with exponents that also depend on θ. Suppose we repeat the updating of the prior-posterior. Then the final posterior will still have the beta-like if-part sets of the form

$$a_{F,s}(\theta) = \nu_{F,s}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{h,k}+\sum_i \alpha_{g,i}\lambda_{ik}(\theta)}\Big(1 - \Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)\Big)^{\beta_{h,k}+\sum_i \beta_{g,i}\gamma_{ik}(\theta)} \tag{101}$$

for $\theta \in D = \cap_i (m_{g,i}, m_{g,i} + d_{g,i}) \cap (m_{h,k}, m_{h,k} + d_{h,k})$.

Corollary 4.3 *Semi-conjugacy of gamma if-part sets.*

Suppose that the SAM-based prior $H(\theta)$ uses gamma (or Poisson) if-part sets $a_{h,k}(\theta) = G(m_{h,k}, d_{h,k}, \alpha_{h,k}, \beta_{h,k}, \nu_{h,k}; \theta)$ and the SAM-based likelihood $G(x|\theta)$ also uses gamma (or Poisson) if-part sets $a_{g,j}(\theta) = G(m_{g,j}, d_{g,j}, \alpha_{g,j}, \beta_{g,j}, \nu_{g,j}; \theta)$ where

$$G(m, d, \alpha, \beta, \nu; \theta) = \nu\Big(\frac{\theta - m}{d}\Big)^{\alpha} e^{-(\frac{\theta-m}{d})/\beta} \tag{102}$$

if $\frac{\theta-m}{d} > 0$ (or if $\theta > m$) for some constant $\nu > 0$. Then the posterior $F(\theta|x)$ in (37) will have set functions $a_{F,i}(\theta)$ of semi-gamma form

$$a_{F,i}(\theta) = a_{g,j}(\theta)\, a_{h,k}(\theta) \tag{103}$$

$$= \nu_{F,i}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{h,k}} e^{-(\frac{\theta-m_{h,k}}{d_{h,k}})/\beta_{h,k}}\Big(\frac{\theta - m_{g,j}}{d_{g,j}}\Big)^{\alpha_{g,j}} e^{-(\frac{\theta-m_{g,j}}{d_{g,j}})/\beta_{g,j}} \tag{104}$$

$$= \nu_{F,i}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{h,k}+\alpha_{g,j}\log_{(\frac{\theta-m_{h,k}}{d_{h,k}})}(\frac{\theta-m_{g,j}}{d_{g,j}})} e^{-(\frac{\theta-m_{h,k}}{d_{h,k}})/\beta_{h,k}-(\frac{\theta-m_{g,j}}{d_{g,j}})/\beta_{g,j}} \tag{105}$$

$$= \nu_{F,i}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{h,k}+\alpha_{g,j}\log_{(\frac{\theta-m_{h,k}}{d_{h,k}})}(\frac{\theta-m_{g,j}}{d_{g,j}})} \times e^{-(\theta-\frac{\beta_{g,j}d_{g,j}m_{h,k}+\beta_{h,k}d_{h,k}m_{g,j}}{\beta_{g,j}d_{g,j}+\beta_{h,k}d_{h,k}})/\frac{\beta_{g,j}\beta_{h,k}d_{g,j}d_{h,k}}{\beta_{g,j}d_{g,j}+\beta_{h,k}d_{h,k}}} \tag{106}$$

if $\theta > m_{h,k}$ and $\theta > m_{g,j}$ (or $\theta > \max\{m_{h,k}, m_{g,j}\}$).

A special case occurs if $m_{h,k} = m_{g,j}$ and $d_{h,k} = d_{g,j}$. Then $a_{F,i}$ has gamma form

$$a_{F,i}(\theta) = \nu_{F,i}\Big(\frac{\theta - m_{h,k}}{d_{h,k}}\Big)^{\alpha_{F,i}} e^{-(\frac{\theta-m_{h,k}}{d_{h,k}})/\beta_{F,i}} = G(m_{h,k}, d_{h,k}, \alpha_{F,i}, \beta_{F,i}, \nu_{F,i}; \theta) \tag{107}$$

if $\theta > m_{h,k}$. Here $\alpha_{F,i} = \alpha_{h,k} + \alpha_{g,j}$, $\beta_{F,i} = \frac{\beta_{g,j}\beta_{h,k}}{\beta_{g,j}+\beta_{h,k}}$, and $\nu_{F,i} = \nu_{h,k}\nu_{g,j}$.

Corollary 4.4 *Semi-conjugacy of Laplace if-part sets.*

Suppose that the SAM-based prior $H(\theta)$ uses Laplace if-part sets $a_{h,k}(\theta) = L(m_{h,k}, d_{h,k}; \theta)$ and the SAM-based likelihood $G(x|\theta)$ also uses Laplace if-part sets

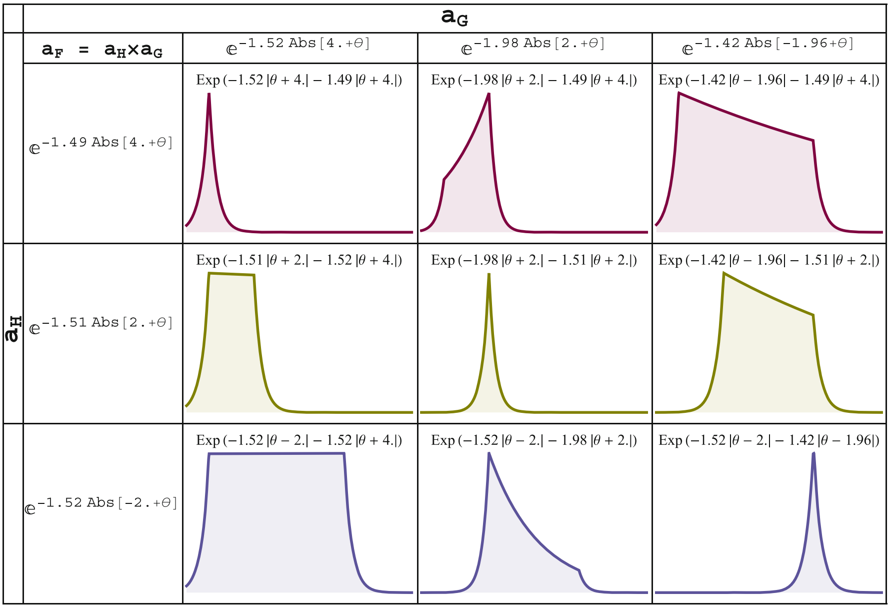

Fig. 6 Laplace semi-conjugacy: *plots* show examples of semi-conjugate Laplace set functions for the doubly fuzzy posterior approximator $F(\theta|x) \propto H(\theta)G(x|\theta)$. The approximator uses five Laplace set functions a_H for the prior approximator $H(\theta)$ and five Laplace set functions a_G for the likelihood approximator $G(x|\theta)$. Thus $F(\theta|x)$ is a weighted sum of 25 Laplacian semi-conjugate set functions of the form $a_{F,i}(\theta) = \exp\left(-|\frac{\theta - m_{h,k}}{d_{h,k}}| - |\frac{\theta - m_{g,j}}{d_{g,j}}|\right)$. The *plots* show that the Laplace semi-conjugate function can have a variety of shapes depending on the location and dispersion parameters of the prior and likelihood set functions

$a_{g,j}(\theta) = L(m_{g,j}, d_{g,j}; \theta)$ where

$$L(m, d; \theta) = e^{-|\frac{\theta - m}{d}|}. \tag{108}$$

Then $F(\theta|x)$ in (37) will have set functions $a_{F,i}(\theta)$ of the (semi-Laplace) form

$$a_{F,i}(\theta) = a_{g,j}(\theta)\, a_{h,k}(\theta) \;=\; e^{-|\frac{\theta - m_{h,k}}{d_{h,k}}| - |\frac{\theta - m_{g,j}}{d_{g,j}}|}. \tag{109}$$

Figure 6 shows examples of these semi-Laplace forms and their shapes. Convex combinations of semi-Laplace set functions give the doubly fuzzy Laplace-SAM posterior approximators in Fig. 7.

A special case occurs if $m_{h,k} = m_{g,j}$ and $d_{h,k} = d_{g,j}$. Then $a_{F,i}$ is of Laplace form

$$a_{F,i}(\theta) = e^{-|\frac{\theta - m_{h,k}}{d_{h,k}/2}|}. \tag{110}$$

Such semi-conjugacy differs from outright conjugacy in a crucial respect: The parameters of semi-conjugate if-part sets increase with each iteration or Bayesian

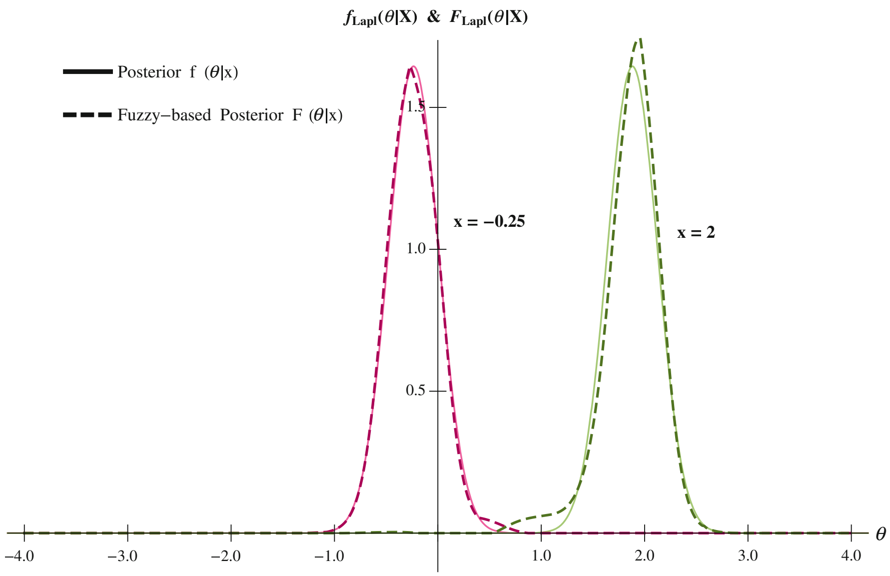

Fig. 7 Doubly Fuzzy Laplace-SAM approximator for two normal posterior pdfs: fuzzy prior and likelihood approximators use Laplacian set functions to generate the posterior approximator $F(\theta|x)$ for the same normal posteriors in Fig. 1. The prior and likelihood fuzzy approximators $H(\theta)$ and $G(x|\theta)$ use Laplace-SAM instead of sinc-SAMs and Gaussian-SAMs. All fuzzy approximators used 15 rules for 6,000 iterations on 500 uniform sample points

update. The conjugate Gaussian sets in Corollary 4.1 avoid this parameter growth while the semi-conjugate beta, gamma, and Laplace sets in Corollaries 4.2–4.4 incur it. The latter if-part sets do not depend on a fixed number of parameters such as centers and widths as in the Gaussian case. Only the set functions with the same centers m_j and widths d_j (in the special cases) will result in set functions for posterior approximation with the same fixed number of parameters. Coping with this form of "parameter explosion" remains an open area of research in the use of fuzzy systems in iterative Bayesian inference.

5 Conclusion

We have shown that additive fuzzy systems can uniformly approximate a Bayesian posterior even in the hierarchical case when the prior pdf depends on its own uncertain parameter with its own hyperprior. This gives a triply fuzzy uniform function approximation. That hyperprior can in turn have its own hyperprior. The result will be a quadruply fuzzy uniform function approximation and so on. This new theorem substantially extends the choice of priors and hyperpriors from well-known closed-form pdfs that obey conjugacy relations to arbitrary rule-based priors that depend on user knowledge or sample data. An open research problem is whether semi-conjugate rules or other techniques can reduce the exponential rule explosion that both doubly and triply fuzzy Bayesian systems face in general Bayesian iterative inference.

Appendix: Proof of Theorem 2

This appendix restates the Extended Bayesian Approximation Theorem and gives its proof.

Theorem 2 *Extended Bayesian Approximation Theorem. Suppose that* $h(\theta|\tau)$, $\pi(\tau)$, *and* $g(x|\theta)$ *are bounded and continuous. Suppose that* $\Pi(\tau)H(\theta|\tau)G(x|\theta) = P(\theta,\tau)G(x|\theta) \neq 0$ *almost everywhere. Then the triply fuzzy SAM system* $F(\theta,\tau|x) = PG/Q$ *uniformly approximates* $f(\theta,\tau|x)$ *for all* $\epsilon > 0 : |F(\theta,\tau|x) - f(\theta,\tau|x)| < \epsilon$ *for all* x *and all* (θ,τ).

Proof Write the posterior pdf $f(\theta,\tau|x)$ as $f(\theta,\tau|x) = \frac{p(\theta,\tau)g(x|\theta)}{q(x)}$ and its approximator $F(\theta,\tau|x)$ as $F(\theta,\tau|x) = \frac{P(\theta,\tau)G(x|\theta)}{Q(x)}$. The SAM approximations for the prior and likelihood functions are uniform (Kosko 1996). So they have approximation error bounds ϵ_p and ϵ_g that do not depend on x or θ:

$$|\Delta P| < \epsilon_p \quad \text{and} \quad |\Delta G| < \epsilon_g \tag{111}$$

where $\Delta P = P(\theta,\tau) - p(\theta,\tau)$ and $\Delta G = G(x|\theta) - g(x|\theta)$. The posterior error ΔF is

$$\Delta F = F - f = \frac{PG}{Q(x)} - \frac{pg}{q(x)}. \tag{112}$$

Expand PG in terms of the approximation errors to get

$$PG = (\Delta P + p)(\Delta G + g) \tag{113}$$

$$= \Delta P \Delta G + \Delta P g + p\Delta G + pg. \tag{114}$$

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

$$\Delta Q = \int_{\mathcal{D}} (\Delta P \Delta G + \Delta P g + p\Delta G + pg - pg)\, d\tau\, d\theta. \tag{115}$$

So

$$|\Delta Q| \leq \int_{\mathcal{D}} |\Delta P \Delta G + \Delta P g + p\Delta G|\, d\tau\, d\theta \tag{116}$$

$$\leq \int_{\mathcal{D}} \big(|\Delta P||\Delta G| + |\Delta P| g + p|\Delta G|\big)\, d\tau\, d\theta \tag{117}$$

$$< \int_{\mathcal{D}} (\epsilon_p \epsilon_g + \epsilon_p g + p\epsilon_g)\, d\tau\, d\theta \quad \text{by (111)}. \tag{118}$$

Parameter set $\mathcal{D}$ has finite Lebesgue measure $m(\mathcal{D}) = \int_{\mathcal{D}} d\tau\, d\theta < \infty$ because $\mathcal{D}$ is a compact subset of a metric space and thus (Munkres 2000) it is (totally) bounded.

Then the bound on ΔQ becomes

$$|\Delta Q| < m(\mathcal{D})\epsilon_p\epsilon_g + \epsilon_g + \epsilon_p \int_{\mathcal{D}} g(x|\theta)\, d\theta \tag{119}$$

because $\int_{\mathcal{D}} p(\theta, \tau) d\tau\, d\theta = 1$ and g has no dependence on τ.

We now invoke the extreme value theorem (Folland 1999). 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

$$|\Delta Q| < \epsilon_q \tag{120}$$

where we define the error bound ϵ_q as

$$\epsilon_q = m(\mathcal{D})\epsilon_p\epsilon_g + \epsilon_g + \epsilon_p \max_x \left\{ \int_{\mathcal{D}} g(x|\theta) d\theta \right\}. \tag{121}$$

Rewrite the posterior approximation error ΔF as

$$\Delta F = \frac{qPG - Qpg}{qQ} \tag{122}$$

$$= \frac{q(\Delta P \Delta G + \Delta P g + p \Delta G + pg) \; - \; Qpg}{q(q + \Delta Q)} \tag{123}$$

Inequality (120) implies that $-\epsilon_q < \Delta Q < \epsilon_q$ and that $(q - \epsilon_q) < (q + \Delta Q) < (q + \epsilon_q)$. Then (111) gives similar inequalities for ΔP and ΔG. So

$$\frac{q[-\epsilon_p\epsilon_g - \min(g)\epsilon_p - \min(h)\epsilon_g] \; - \; \epsilon_q pg}{q(q - \epsilon_q)} < \Delta F < \frac{q[\epsilon_p\epsilon_g + \max(g)\epsilon_p + \max(h)\epsilon_g] \; + \; \epsilon_q pg}{q(q - \epsilon_q)}. \tag{124}$$

The extreme value theorem ensures that the maxima in (124) 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 (124) guarantees at best a pointwise approximation of $f(\theta, \tau|x)$ 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 pg 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 $|q| \leq \left(\int_{\mathcal{D}} |p|\, d\tau\, d\theta\right) \left(\|g(x, \theta)\|_\infty\right) = \|g(x, \theta)\|_\infty$ since p 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

$$\epsilon_- < \Delta F < \epsilon_+ \tag{125}$$

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

$$\epsilon_- = \frac{\left(-\epsilon_p \epsilon_g - \min\{g\}\epsilon_p - \min\{p\}\epsilon_g\right)\min\{q\} - pg\epsilon_q}{\min\{q\}\left(\min\{q\} - \epsilon_q\right)} \tag{126}$$

$$\epsilon_+ = \frac{\left(\epsilon_p \epsilon_g + \max\{g\}\epsilon_p + \max\{p\}\epsilon_g\right)\max\{g\} + pg\epsilon_q}{\min\{q\}\left(\min\{q\} - \epsilon_q\right)}. \tag{127}$$

Now $\epsilon_q \to 0$ as $\epsilon_g \to 0$ and $\epsilon_p \to 0$. 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 p and g fix or determine q independent of the approximators P and G involved and because $\epsilon_q \to 0$ when $\epsilon_p \to 0$ and $\epsilon_g \to 0$. So we can achieve arbitrarily small ϵ_q that satisfies $\epsilon_q < \min\{q\}$ by choosing appropriate ϵ_p and ϵ_g. Then $\Delta F \to 0$ as $\epsilon_g \to 0$ and $\epsilon_p \to 0$. So $|\Delta F| \to 0$.

Theorem 2 now follows from the following lemma: If Y is compact and $f_n \longrightarrow f$ uniformly then

$$\int_Y f_n(x, y, \mathbf{z})\, dy \longrightarrow \int_Y f(x, y, \mathbf{z})\, dy \quad \text{uniformly.} \tag{128}$$

The result guarantees that uniformity in approximation still holds after marginalizing a multidimensional uniform approximator. This result implies Theorem 2 because we have uniform approximators for $f(\theta, \tau|x)$. We can marginalize over τ to get a posterior approximation in terms of just θ. Thus $F \longrightarrow f$ uniformly implies $\int F\, d\tau \longrightarrow \int f\, d\tau$ uniformly.

We now prove this lemma (128). The uniform convergence of the sequence f_n to f implies that for all $\epsilon > 0$ there is an $n \in \mathbb{N}$ such that

$$|f_n(x, y, \mathbf{z}) - f(x, y, \mathbf{z})| < \epsilon$$

for all $(x, y, \mathbf{z}) \in X \times Y \times \prod Z_i$. Thus

$$-\epsilon < f_n(x, y, \mathbf{z}) - f(x, y, \mathbf{z}) < \epsilon. \tag{129}$$

Thus

$$-\int_Y \epsilon\, dy < \int_Y f_n(x, y, \mathbf{z})\, dy - \int_Y f(x, y, \mathbf{z})\, dy$$

and

$$\int_Y f_n(x, y, \mathbf{z})\, dy - \int_Y f(x, y, \mathbf{z})\, dy \;<\; \int_Y \epsilon\, dy\ .$$

Y has finite Lebesgue measure $m(Y) = \int_Y dy$ because Y is compact. Define $s_n(x, \mathbf{z}) = \int_Y f_n(x, y, \mathbf{z})\, dy$ and $s(x, \mathbf{z}) = \int_Y f(x, y, \mathbf{z})\, dy$. Then

$$-\,\epsilon m(Y) \;<\; s_n(x, \mathbf{z}) - s(x, \mathbf{z}) \;<\; \epsilon m(Y)\ . \tag{130}$$

$$\text{Thus}\quad |s_n(x, \mathbf{z}) - s(x, \mathbf{z})| \;<\; \epsilon m(Y)\ . \tag{131}$$

Define ϵ' as $\epsilon' = \epsilon m(Y)$. Then for all $\epsilon' > 0$ there exists an $n \in \mathbb{N}$ such that

$$|s_n(x, \mathbf{z}) - s(x, \mathbf{z})| \;<\; \epsilon'$$

for all $(x, \mathbf{z}) \in X \times \prod Z_i$.

Therefore

$$\int_Y f_n(x, y, \mathbf{z})\, dy \longrightarrow \int_Y f(x, y, \mathbf{z})\, dy \tag{132}$$

uniformly in x and $\mathbf{z}$. □

References

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical statistics: Volume I* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Billingsley, P. (1995). *Probability and measure* (3rd ed.p. 269). London: Wiley.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York, NY: McGraw-Hill.

Folland, G. B. (1999). *Real analysis: Modern techniques and their applications* (2nd ed.). New York, NY: Wiley-Interscience.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2005). *Introduction to mathematical statistics* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

Jin, Y. (2000). Fuzzy modeling of high-dimensional systems: Complexity reduction and interpretability improvement. *IEEE Transactions on Fuzzy Systems, 8*(2), 212–221.

Kosko, B. (1992). *Neural networks and fuzzy systems: A dynamical systems approach to machine intelligence*. Englewood Cliffs, NJ: Prentice Hall.

Kosko, B. (1994). Fuzzy systems as universal approximators. *IEEE Transactions on Computers, 43*(11), 1329–1333.

Kosko, B. (1995). Optimal fuzzy rules cover extrema. *International Journal of Intelligent Systems, 10*(2), 249–255.

Kosko, B. (1996). *Fuzzy engineering*. Englewood Cliffs, NJ: Prentice Hall.

Kosko, B. (2004). Probable equality, superpower sets, and superconditionals. *International Journal of Intelligent Systems, 19*, 1151–1171.

Lee, I., Kosko, B., & Anderson, W. F. (2005). Modeling gunshot bruises in soft body armor with adaptive fuzzy systems. *IEEE Transactions on Systems, Man, and Cybernetics, 35*(6), 1374–1390.

Mitaim, S., & Kosko, B. (1998). Neural fuzzy agents for profile learning and adaptive object matching. *Presence, 7*(6), 617–637.

Mitaim, S., & Kosko, B. (2001). The shape of fuzzy sets in adaptive function approximation. *IEEE Transactions on Fuzzy Systems, 9*(4), 637–656.

Mitra, S., & Pal, S. K. (1996). Fuzzy self-organization, inferencing, and rule generation. *IEEE Transactions on Systems, Man, Cybernetics—A, 26*(5), 608–620.

Munkres, J. R. (2000). *Topology* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Osoba, O., Mitaim, S., & Kosko, B. (2011). Bayesian inference with adaptive fuzzy priors and likelihoods, *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics* (to appear).

Watkins, F. A.(1994). *Fuzzy engineering*. Ph.D. dissertation, Department of Electrical Engineering, UC Irvine.

Watkins, F. A. (1995) 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).

Xu, R., & Wunsch, D. C. (2009). *Clustering*. New York, NY: IEEE Press, Wiley.