Uniform Convergence of Probability Mixtures that Represent Combined Fuzzy Systems

Bart Kosko
Department of Electrical and Computer Engineering
University of Southern California
Los Angeles, CA 90089-2564
kosko@usc.edu

***Abstract*—We can combine expert knowledge by combining the probability mixtures that represent the if-then rules of the experts. Fuzzy rules define a generalized probability mixture whose moments describe a fuzzy system and its uncertainty. The mixture's Bayesian structure gives a complete posterior probability description of the if-then fuzzy-set rules as they fire. A new theorem extends the uniform convergence of a fuzzy system's mixture to the uniform convergence of the sequence of expert mixtures that represent any number of combined fuzzy systems as they each converge to a target function. A mixture of just two normal bell curves exactly represents the target function in the scalar case and serves as the probabilistic target of the converging mixture sequence. A sampled deep neural network can serve as the target function. Then the mixture defines a proxy system that gives a probabilistic form of explainable AI. The uniform convergence result extends to any continuous transformation of the converging fuzzy systems and further extends to the uniform mixture convergence of any continuous function of the combined systems and their continuous transformations.**

***Index Terms*—probability mixtures, fuzzy systems, XAI, combined systems, Bayesian rule posteriors, uniform convergence**

I. Mixture Representations of Combined Fuzzy Systems for *Probabilistic XAI*

A core epistemic problem of artificial intelligence is knowledge combination [1]–[7]: *How do we combine the knowledge of experts?* Do we just combine or average what the experts say? Or can we somehow combine what they *know*?

This paper models combining what experts know by combining the express or implied if-then fuzzy rules that approximate the input-output behavior of the experts. We then prove that generalized probability mixtures represent combined fuzzy systems: The mixtures converge uniformly to the mixture that represents some target function when each of the combined fuzzy system converges to that target function. This result gives a complete probabilistic description of the fuzzy systems and of the combined system. The target function can be a sampled trained neural network or other function approximator or a sampled closed-form equation.

Figures 1 and 2 display some of this probabilistic description for lone and combined fuzzy systems. Figure 1 shows the three probability mixture surfaces $p^1(y|x)$, $p^2(y|x)$, and $p^3(y|x)$ of the respective fuzzy systems F^1, F^2, and F^3 after each adaptive fuzzy system F_n^k has converged to the same sampled target function $f(x) = \sin(x)$. Fuzzy system F^1 contains 10 if-then rules $R_{A_1^1 \to B_1^1}, \ldots, R_{A_{10}^1 \to B_{10}^1}$ with Gaussian if-part fuzzy sets $A_j^1 \subset R^n$. Fuzzy system F^2 contains 15 Gaussian rules $R_{A_j^2 \to B_j^2}$. Fuzzy system F^3 contains 10 sinc rules $R_{A_j^3 \to B_j^3}$ where the 10 if-part sets A_j^3 have *sinc* or Shannon-wavelet form $\text{sinc}(x) = \frac{sin(x)}{x}$. The fourth mixture $p(y|x)$ corresponds to the converged *combined* fuzzy system F that combines F^1, F^2, and F^3 by combining their throughputs or rules rather than by just averaging their outputs as with random forests [8]–[10]. Convergence of the lone systems F_n^k to the target f drives the convergence of their mixtures $p_n^k(y|x)$ to $p^k(y|x)$. This drives the convergence of the master mixture $p_n(y|x)$ of the combined system F_n.

Figure 2 shows some of the Bayesian rule and subsystem posteriors for the 3 fuzzy systems F^1, F^2, and F^3 and their combined system F. The posteriors $p^k(j|x, y)$ describe the relative firings of all rules for each input x and observed output y. A higher-level posterior $p(k|x, y)$ describes the relative importance of each fuzzy subsystem F^k to the combined fuzzy system F for each input x.

These mixture-based probability descriptions give a new form of explainable AI (XAI) [11]–[14] that we here call *probabilistic* XAI. The description includes Bayesian posterior probabilities over both the combined fuzzy systems F^k and over their rules and over their conditional variances and other higher moments. This information can help prune some of the combined fuzzy systems or their rules and may suggest other systems to add to the combination [15]. Combining fuzzy systems can improve classification performance [16] just as combining randomly generated trees can reduce variance and improve other statistical measures [9]. Simulations likewise showed that combined fuzzy systems tended to give better performance than did their lone subsystems.

II. How Combined Fuzzy Systems Define Generalized Probability Mixtures

Suppose the fuzzy system $F : R^n \to R^p$ combines q fuzzy systems $F^1, \ldots, F^q$. The simplest way to combine the systems is to combine their outputs through a fixed-weight average: $F = \frac{1}{n} \sum_{k=1}^{1} F^k$. This average is a special case of a generalized convex combination of subsystem outputs:

$$F(x) = \sum_{k=1}^{q} \lambda^k(x) F^k(x) \tag{1}$$

978-1-6654-6710-0/22/$31.00 ©2022 IEEE

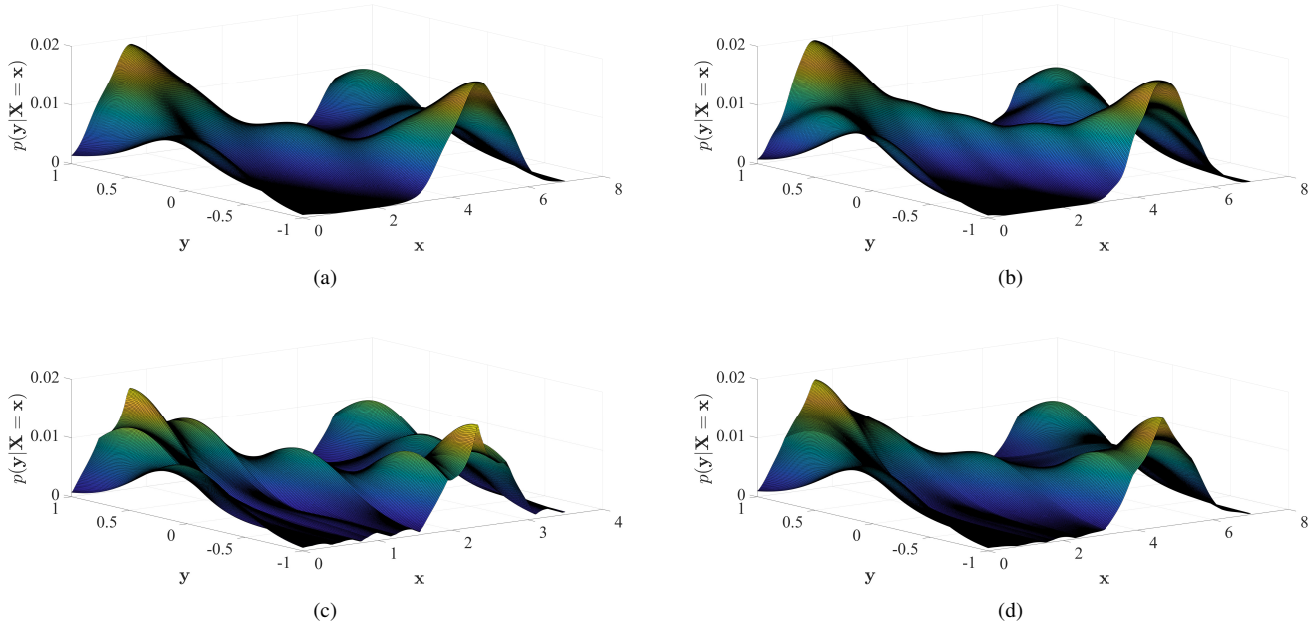

Fig. 1: *Probability mixture* $p^k(y|x) = p_1^k(x)\ p_{B_1^k}(y|x) + \cdots + p_{m_k}^k(x)\ p_{B_{m_k}^k}(y|x)$ *surfaces of 3 converged additive fuzzy systems* F^1, F^2, *and* F^3 *and the converged mixture surface of the fuzzy system* F *that combined these three fuzzy systems with throughput combination. The adaptive fuzzy systems* F_n^k *and* F_n *converged quickly to the same sampled target function* $f(x) = \sin(x)$ *on a* 2π *segment of its domain. The average of each mixture surface gives back the target* f*:* $E[Y|X = x] = \sin(x)$*. (a) Mixture* $p^1(y|x)$ *of the converged 10-rule Gaussian additive fuzzy system* F^1*. (b) Mixture* $p^2(y|x)$ *of the converged 15-rule Gaussian additive fuzzy system* F^2*. (c) Mixture* $p^3(y|x)$ *of the converged 10-rule sinc fuzzy system* F^3 *where the scalar if-part fuzzy sets* A_j^3 *have the wavelet form* $\text{sinc}(x) = \frac{\sin(x)}{x}$*. (d) Mixture* $p(y|x)$ *of the throughput-combined fuzzy system* F *that combined* F^1*,* F^2*, and* F^3 *by combining their throughputs or rules rather than by averaging their outputs.*

for generalized convex coefficients λ^k that depend on the input vector x: $\lambda^k(x) \geq 0$ and $\lambda^1(x) + \cdots + \lambda^q(x) = 1$ for all x. The arithmetic mean results when $\lambda^k(x) = \frac{1}{n}$ for all k and all x. A more powerful way to combine fuzzy systems combines their *throughputs* or rules before averaging per (33) - (43). The theorems below apply to both types of combination. Figures 1 and 2 show a fuzzy system F that uses throuput combination to combine the additive fuzzy systems F^1, F^2, and F^3.

The kth additive fuzzy system F^k combines m_k rules $R_{A_1^k \to B_1^k}, \ldots, R_{A_{m_k}^k \to B_{m_k}^k}$ through some *knowledge combination* operator $\mathcal{C}^k$: $F^k = \mathcal{C}^k(R_{A_1^k \to B_1^k}, \ldots, R_{A_{m_k}^k \to B_{m_k}^k})$. We use an additive $\mathcal{C}^k$ throughout. The jth if-then rule $R_{A_j^k \to B_j^k}$ of F^k associates the jth if-part fuzzy set $A_j^k \subset R^n$ to the jth then-part fuzzy set $B_j^k \subset R^p$. The associated sets A_j^k and B_j^k are fuzzy [17], [18] because their respective set indicator functions a_j^k and b_j^k are multivalued: $a_j^k : R^n \to [0, 1]$ and $b_j^k : R^p \to [0, 1]$.

The system F^k adds the *fired* rule outputs $b_j^k(y|x)$ for input x. It then computes its output $F^k(x)$ as the centroid or other average of these summed rule firings: $F^k(x) = \text{Centroid}(B^k(y|x))$. The fired rule sum $b^k(y|x) = w_1^k b_1^k(y|x) + \cdots + w_{m_k}^k b_{m_k}^k(y|x)$ uses the nonnegative rule weights $w_j^k \geq 0$ to weight the respective fired rule then-parts $b_j^k(y|x)$. It thereby weights the system's stored jth rule $R_{A_j^k \to B_j^k}$ itself: $R_{A_j^k \to B_j^k}(x) = w_j^k a_j^k(x) B_j^k$ and so $r_{A_j^k \to B_j^k}(x) = w_j^k b_j^k(y|x) = w_j^k a_j^k(x) b_j^k(y)$. Vector input x_0 formally fires the jth rule $R_{A_j^k \to B_j^k}$ when x_0 as the vector delta spike $\delta(x - x_0)$ convolves with the rule function $r_{A_j^k \to B_j^k}$: $b_j(y|x_0) = \int_{\mathbb{R}^n} \delta(x - x_0)\ r_{A_j^k \to B_j^k}(x, y)\ dx = \int_{\mathbb{R}^n} \delta(x - x_0)\ a_j^k(x)\ b_j^k(y)\ dx = b_j^k(y) \int_{\mathbb{R}^n} \delta(x - x_0)\ a_j^k(x)\ dx = b_j^k(y)\ a_j^k(x_0)$ by the sifting of the delta function. Almost all fuzzy systems in practice have this additive form [19]–[22].

The fuzzy system F^k gives rise to a generalized probability mixture $p^k(y|x)$ so long as the combined fired then-part sets are integrable and not negative. The mixture $p^k(y|x)$ mixes m_k generalized priors $p_j^k(x)$ with m_k likelihoods $p_{B_j}^k(y|x)$:

$$p^k(y|x) = p_1^k(x)\ p_{B_1^k}(y|x) + \cdots + p_{m_k}^k(x)\ p_{B_{m_k}^k}(y|x). \quad (2)$$

The jth rule $R_{A_j^k \to B_j^k}$ of the additive fuzzy system F^k corresponds to the jth mixed term $p_j^k(x) p_{B_j^k}(y|x)$ in the sum. The m_k prior probabilities $p_j^k(x)$ are themselves convex mixing weights for each input x.

The mixture result (2) follows from additivity and the assumption that $b^k(y|x) \geq 0$ and not trivially equal to zero and that its integral is finite. Then $p^k(y|x) = \frac{b^k(y|x)}{\int b^k(y|x) dy}$ is a probability density function since $p^k(y|x) \geq 0$ and $p^k(y|x)$ integrates to unity. The additive rule firings $b^k(y|x) = w_1^k b_1^k(y|x) + \cdots + w_{m_k}^k b_{m_k}^k(y|x)$ gives the result for the *standard*-additive rule firing $b_j^k(y|x) = a_j^k(x) b_j^k(y)$ if $a_j^k(x) >$

0 with then-part volume $V_j^k = \int b_j^k(y)dy$: $p^k(y|x) = \sum_{j=1}^{m_k} \frac{w_j^k a_j^k(x) V_j^k}{\sum_{i=1}^{m_k} w_i^k a_i^k(x) V_i^k} \frac{b_j^k(y)}{V_j^k} = \sum_{j=1}^{m_k} p_j^k(x) p_{B_j}^k(y)$ with priors $p_j^k(x) = \frac{w_j^k a_j^k(x) V_j^k}{\sum_{i=1}^{m_k} w_i^k a_i^k(x) V_i^k}$ and likelihoods $p_{B_j}^k(y) = \frac{b_j^k(y)}{V_j^k}$. Putting this together gives the mixture result in (2):

$$p^k(y|x) = \frac{b^k(y|x)}{\int b^k(y|x)dy} = \frac{\sum_{j=1}^{m_k} w_j^k b_j^k(y|x)}{\sum_{i=1}^{m_k} w_i^k \int b_i^k(y|x)dy} \tag{3}$$

$$= \sum_{j=1}^{m_k} \frac{w_j^k a_j^k(x) V_j^k}{\sum_{i=1}^{m_k} w_i^k a_i^k(x) V_i^k} \frac{b_j^k(y)}{V_j^k} = \sum_{j=1}^{m_k} p_j^k(x) p_{B_j^k}(y|x) \tag{4}$$

The mixture sum (2) itself just states a generalized version of the elementary theorem on total probability. So it implies a rule-based Bayes theorem as a corollary. Suppose that vector input x passes through fuzzy system F^k and produces the output scalar or vector y: $y = F^k(x)$. Then the posterior probability $p^k(j|x,y) = P(R_{A_j^k \to B_j^k} | X = x, Y = y)$ gives the probability or degree to which the kth fuzzy system's jth rule $R_{A_j^k \to B_j^k}$ fires given the input x and output y:

$$p^k(j|x,y) = \frac{p_j^k(x)\ p_{B_j^k}(y|x)}{p^k(y|x)} = \frac{p_j^k(x)\ p_{B_j^k}(y|x)}{\sum_{l=1}^{m} p_l^k(x)\ p_{B_l^k}(y|x)}. \tag{5}$$

The Bayesian posterior (5) is a natural if unforeseen benefit of the mixture structure. It describes a fuzzy system's "gray box" of parallel rules in a quantitative and principled way because it gives an input-by-input and rule-by-rule description of how the fuzzy system's rule ensemble maps inputs to outputs. Simulations confirmed what Figure 2 shows: Large-rule fuzzy systems tend to fire only a small number of rules to a nontrivial degree for a given input x.

The expectation or first moment of the mixture $p^k(y|x)$ gives back the system F^k itself:

$$F^k(x) = E_{p^k(y|x)}[Y|X = x] = \int y\ p^k(y|x)dy. \tag{6}$$

This first-moment result follows from the structure of the mixture $p^k(y|x)$ and so thereby avoids the earlier ad hoc constructions of fuzzy systems as "fuzzifiers" that use output centroids or other aggregation operations. The second moment gives the system's conditional variance $V^k[Y|X = x]$ as the convex sum of then-part and interpolation uncertainties [23]:

$$V^k[Y|X = x] = \sum_{j=1}^{m_k} p_j^k(x)\ \sigma_{B_j^k}^2 + \sum_{j=1}^{m_k} p_j^k(x)\ [c_j^k - F^k(x)]^2 \tag{7}$$

for then-part set variance $\sigma_{B_j^k}^2 = \int (y - c_j^k)^2 p_{B_j^k}(y)\ dy$ where c_j^k is the centroid or raw first moment of the system F^k's jth then-part set B_j^k: $c_j^k = \frac{\int y\ p_{B_j^k}(y)dy}{V_j^k}$.

Additive fuzzy systems are also universal function approximators on compact sets [24], [25] just as are feedforward neural networks with sigmoidal hidden units [26]–[28]. So they can serve as proxy systems for sampled black boxes such as trained neural classifiers or regressors if the user can control the fuzzy system's inherent exponential rule explosion and if the learning laws are practical given the training samples taken from the target black box.

Rule explosion is endemic because the rules form a graph cover of the approximated target system f in the input-output product space. Mixtures can ameliorate this problem in some cases if the system draws its rules at random from a properly trained mixture $p(y|x)$ and thus if the system samples from a virtual rule continuum for each input x. Then weighted Monte Carlo can approximate the output $F(x)$ because it is an expectation [29].

III. Uniform Convergence of Mixtures that Represent *OUTPUT*-Combined Fuzzy Systems

We start with the important special case of probability mixtures that represent *output*-combined fuzzy systems F.

Suppose that each of the q fuzzy systems F_n^k converges *uniformly* to the target function f. This convergence can correspond to neural-like supervised or unsupervised learning at discrete iterations n. So n indexes the family F_n^k of fuzzy systems as it converges to f. Then for all $\epsilon > 0$ there exists a positive integer n_0^k such that for all $n \geq n_0^k$: $|F_n^k(x) - f(x)| < \epsilon$ for all x. Uniform convergence entails that the value n_0^k does not depend on any given input value x. This contrasts with *pointwise* convergence that does so depend.

We first prove as Lemma 1 that a generalized convex sum of outputs $F_n(x) = \sum_{k=1}^{q} \lambda^k(x) F_n^k(x)$ converges uniformly to the same target function f for any such set of q generalized convex coefficients $\lambda^1(x), \ldots, \lambda^q(x)$.

Lemma 1: *Uniform convergence of generalized convex sums: The convexly combined system F_n converges uniformly to the target function f if $F_n(x) = \sum_{k=1}^{q} \lambda^k(x) F_n^k(x)$ and if the q subsystems $F_n^1, \ldots, F_n^q$ converge uniformly to the same target function f for any generalized convex coefficients $\lambda^k(x) \geq 0$: $\sum_{k=1}^{q} \lambda^k(x) = 1$ for each x.*

Proof : Suppose that subsystem F_n^k converges uniformly to the target function f. Pick any small $\epsilon > 0$. Then there is some positive integer n_0^k such that for all $n \geq n_0^k$: $|F_n^k(x) - f(x)| < \epsilon$. Define $n_0 = \max\{n_0^1, \ldots, n_0^q\}$. Then for *all* k and for all $n \geq n_0$: $|F_n^k(x) - f(x)| < \epsilon$ for all x. So

$$|F_n(x) - f(x)| = |\sum_{k=1}^{q} \lambda^k(x) F_n^k(x) - \sum_{k=1}^{q} \lambda^k(x) f(x)| \tag{8}$$

$$\leq \sum_{k=1}^{q} \lambda^k(x) |F_n^k(x) - f(x)| \tag{9}$$

$$< \sum_{k=1}^{q} \lambda^k(x) \epsilon = \epsilon \sum_{k=1}^{q} \lambda^k(x) = \epsilon \tag{10}$$

for all x by the triangle inequality and because the q weights $\lambda^k(x)$ are convex and so $\lambda^k(x) \geq 0$ and $\lambda^1(x) + \cdots + \lambda^q(x) = 1$. So F_n converges uniformly to f. $\mathbf{Q.E.D.}$

The same argument shows that F_n converges uniformly to the *mixed* target functions $\sum_{k=1}^{q} \lambda^k f^k$ if each subsystem F_n^k converges uniformly to a distinct target function f^k.

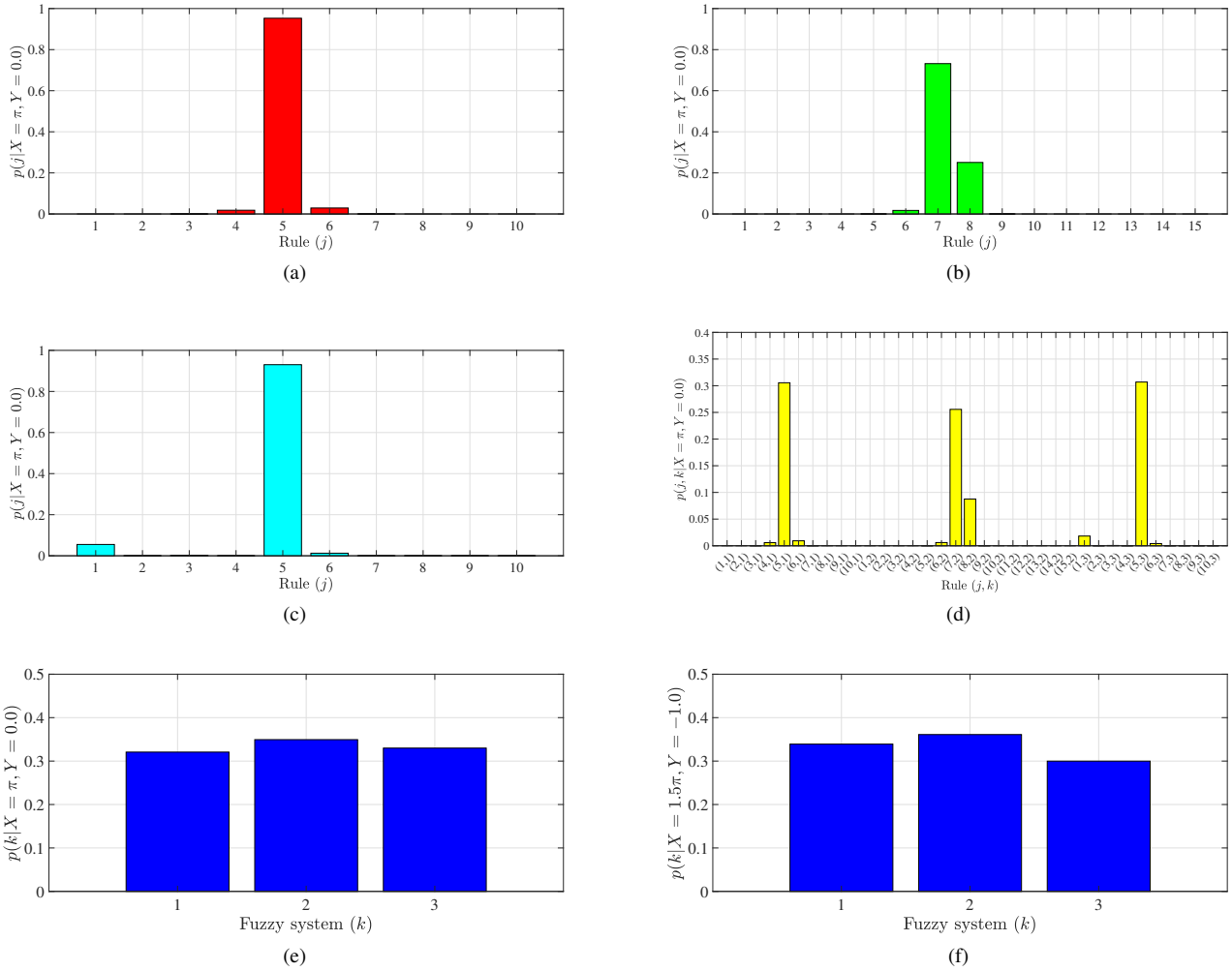

Fig. 2: *Bayesian posteriors over the rules of the 3 separate additive fuzzy systems F^k in Figure 1, over the 35 rules of the throughput-combined fuzzy system F, and over its 3 combined fuzzy subsystems. Each converged fuzzy system closely approximated the sampled target function* $\sin(x)$. *(a) Rule posterior $p^1(j\text{th rule } R_{A_j^1 \to B_j^1}|X=\pi, Y=0)$ for the 10-rule Gaussian fuzzy system F^1 in Figure 1(a) when the input $X=x=\pi$ produced the observed output $F^1(\pi)=0$. The 5th rule $R_{A_5^1 \to B_5^1}$ fired by far to the highest degree when the input was π. (b) Rule posterior $p^2(j|X=\pi, Y=0)$ for the 15-rule Gaussian fuzzy system F^2 in Figure 1(b). The 7th and 8th rules $R_{A_7^2 \to B_7^2}$ and $R_{A_8^2 \to B_8^2}$ fired to the highest degree. (c) Rule posterior $p^3(j|X=\pi, Y=0)$ for the 10-rule sinc fuzzy system F^3 in Figure 1(c). The 5th rule $R_{A_5^3 \to B_5^3}$ fired by far to the highest degree. (d) Rule posterior $p(j,k|X=\pi, Y=0)$ for the 35 rules of the throughput combined system F in Figure 1(d). (e) Subsystem posterior $p(k|X=\pi, Y=0)$ of F over its 3 subsystems F^1, F^2, and F^3 when input $X=x=\pi$ produced the output $F(\pi)=0$. The 15-rule Gaussian-SAM subsystem F^2 contributed slightly more than did the other two subsystems F^1 and F^3 in this case. (f) Subsystem posterior $p(k|X=1.5\pi, Y=1)$ of F when input $X=x=1.5\pi$ produced $F(1.5\pi)=1$.*

The next convergence result is more complicated. It states that the 2-bell-curve mixtures $q_n(y|x)$ that exactly represent the combined fuzzy systems F_n converge uniformly to the 2-bell-curve mixture $p(y|x)$ that exactly represents f if the underlying fuzzy systems F_n converge uniformly to f.

This convergence result requires a basic inheritance fact on uniform boundedness: A uniformly convergent sequence of bounded functions is uniformly bounded. The fact follows because we assume throughout that each fuzzy system F_n^k is individually bounded for each k and each n as F_n^k converges uniformly to the target function f. The function F_n^k is bounded if there is a $B_n^k > 0$ such that $|F_n^k(x)| < B_n^k$ for all x. This will also imply that the target function f is itself bounded. Suppose that F_n^k converges uniformly to f. So for all $n \geq n_0$: $|F_n^k(x) - f(x)| < \frac{\epsilon}{2}$ for all x. Then $|F_n^k(x)| - |F_{n_0}^k(x)| \leq ||F_n^k(x)| - |F_{n_0}^k(x)|| \leq |F_n^k(x) - F_{n_0}^k(x)| \leq |F_n^k(x) - f(x)| + |f(x) - F_{n_0}^k(x)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. This bounds the tail: $|F_n^k(x)| < \epsilon + |F_{n_0}^k(x)| < \epsilon + B_{n_0}^k$ for all x and all $n \geq n_0$. Put $B = 1 + \max\{B_1^k, \ldots, B_{n_0}^k, \epsilon + B_{n_0}^k\}$. Then $|F_n^k(x)| < B$ for all $n \geq 1$ and all x. So B *uniformly bounds* the fuzzy-system sequence F_n^k. It also bounds the target function f because $|f(x)| \leq |f(x) - F_n^k(x)| + |F_n^k(x)| < \frac{\epsilon}{2} + B$.

We use the recent exact mixture representation of any bounded function $f : R^n \to R$ as a convex combination of just 2 normal bell curves [23] centered at the infimum α and the supremum of β in the general case when f is not constant so that $\alpha < \beta$:

$$p(y|x) = w(x)N_\alpha(y|\alpha, \sigma_\alpha^2) + (1 - w(x))N_\beta(y|\beta, \sigma_\beta^2) \quad (11)$$

for *any* variances $\sigma_\alpha^2 > 0$ and $\sigma_\beta^2 > 0$. The generalized convex weights $w(x)$ and $1 - w(x)$ are *Watkins coefficients* [30]:

$$w(x) = \frac{\beta - f(x)}{\beta - \alpha} \quad (12)$$

and thus $1 - w(x) = \frac{f(x)-\alpha}{\beta-\alpha}$. Then the mixture average $E_p[Y|X = x]$ exactly represents the bounded function f: $E_p[Y|X = x] = \int y\ p(y|x)\ dy = (\frac{\beta-f(x)}{\beta-\alpha})\alpha + (\frac{f(x)-\alpha}{\beta-\alpha})\beta = f(x)$ since integrating y against the normal bell curves gives back the respective modes or centroids α and β. The two mixed bell curves collapse to delta spikes at α and β when their variances σ_α^2 and σ_β^2 shrink to zero. This limiting case gives back the original Watkins deterministic two-rule representation of f [30].

Define the 2-bell-curve mixture $q_n^k(y|x)$ of the kth fuzzy system F_n^k at iteration n of its convergence or learning:

$$q_n^k(y|x) = v_n^k(x)n_{\alpha^k}(y) + (1 - v_n^k(x))n_{\beta^k}(y) \quad (13)$$

where we write $n_{\alpha^k}(y) = N_{\alpha^k}(y|\alpha^k, \sigma_{\alpha^k}^2)$ and $n_{\beta^k}(y) = N_{\beta^k}(y|\beta^k, \sigma_{\beta^k}^2)$ for the two normal probability densities. The mixtures use the Watkins coefficients $v_n^k(x) = \frac{\beta^k - F_n^k(x)}{\beta^k - \alpha^k}$ and $1 - v_n^k(x) = \frac{F_n^k(x) - \alpha^k}{\beta^k - \alpha^k}$. The uniform boundedness of the sequence $\{F_n^k\}$ allows us to write $\alpha^k \leq F_n^k(x) \leq \beta^k$ for all x. Again we assume that the fuzzy systems are not constant and so $\alpha^k < \beta^k$. Then $F_n^k(x) = E_{q_n^k}[Y|X = x]$.

Define likewise the 2-bell-curve mixture $q_n(y|x)$ of the combined fuzzy system F_n as

$$q_n(y|x) = v_n(x)n_\alpha(y) + (1 - v_n(x))n_\beta(y) \quad (14)$$

where $\alpha = \min\{\alpha^1, \ldots, \alpha^q\}$ and $\beta = \max\{\beta^1, \ldots, \beta^q\}$ and so $\alpha < \beta$. The Watkins mixing coefficients or priors are $v_n(x) = \frac{\beta - F_n(x)}{\beta - \alpha}$ and $1 - v_n(x) = \frac{F_n(x) - \alpha}{\beta - \alpha}$.

Lemma 2 states that the 2-bell-curve Gaussian mixtures $q_n(y|x)$ are convex combinations of the individual 2-bell-curve Gaussian mixtures $q_n^k(y|x)$ if F_n convexly combines the q fuzzy systems F_n^k with the same generalized convex coefficients $\lambda^k(x)$ for each x at iteration n.

Lemma 2: *Convexity of Gaussian mixtures for mixed systems:*

$$q_n(y|x) = \sum_{k=1}^{q} \lambda^k(x)\ q_n^k(y|x) \quad (15)$$

for any generalized convex coefficients $\lambda^k(x)$ and for any x if the system F_n convexly combines q bounded nonconstant systems F_n^k: $F_n(x) = \sum_{k=1}^q \lambda^k(x)\ F_n^k(x)$.

Proof : Use the 2-bell-curve mixtures (14) and (13) and the convex combination $F_n(x) = \sum_{k=1}^q \lambda^k(x)\ F_n^k(x)$ to give

$$q_n(y|x) = v_n(x)n_\alpha(y) + (1 - v_n(x))n_\beta(y) \quad (16)$$

$$= \frac{\beta - F_n(x)}{\beta - \alpha} n_\alpha(y) + \frac{F_n(x) - \alpha}{\beta - \alpha} n_\beta(y) \quad (17)$$

$$= \frac{1}{\beta - \alpha}[\sum_{k=1}^{q} \lambda^k(x)(\beta - F_n^k(x))n_\alpha(y) + \sum_{k=1}^{q} \lambda^k(x)(F_n^k(x) - \alpha)n_\beta(y)] \quad (18)$$

$$= \sum_{k=1}^{q} \lambda^k(x)\ [v_n^k(x)n_{\alpha^k}(y) + (1 - v_n^k(x))n_{\beta^k}(y)] \quad (19)$$

$$= \sum_{k=1}^{q} \lambda^k(x)\ q_n^k(y|x) \quad (20)$$

where $\alpha = \min\{\alpha^1, \ldots, \alpha^q\}$ and $\beta = \max\{\beta^1, \ldots, \beta^q\}$ and so $\alpha < \beta$. $\mathbf{Q.E.D.}$

Theorem 1 states the main theorem on the uniform convergence of the 2-bell-curve mixture $q_n(y|x)$ that represents the combined fuzzy system F_n as it converges to the 2-bell-curve mixture $p(y|x)$ that represents the target function f if each convex-combined bounded fuzzy system F_n^k converges uniformly to f. The proof uses a bound on the mixed normal likelihoods $n_\alpha(y)$ and $n_\beta(y)$ in terms of their respective mode values $n_\alpha(\alpha)$ and $n_\beta(\beta)$:

$$|n_\alpha(y) - n_\beta(y)| \leq \max(n_\alpha(\alpha), n_\beta(\beta)) \quad (21)$$

$$< \max(n_\alpha(\alpha), n_\beta(\beta)) + 1 \equiv D \quad (22)$$

for all y.

Theorem 1: Uniform Mixture Convergence of Convexly Combined Bounded Fuzzy Systems.

The mixture sequence $q_n(y|x)$ in (14) of the combined fuzzy system F_n converges uniformly to the 2-bell-curve mixture $p(y|x)$ in (11) of the target function f if each fuzzy system F_n^k converges uniformly to f for any generalized convex coefficients $\lambda^k(x)$ for any x if the combined system F_n is a convex combination of q bounded nonconstant systems F_n^k: $F_n(x) = \sum_{k=1}^q \lambda^k(x)\ F_n^k(x)$.

Proof : Suppose that the bounded fuzzy system F_n^k converges uniformly to f for each k. Then the target function f inherits the boundedness of F_n^k systems as shown above because the convergence is uniform.

The uniform convergence of subsystem F_n^k means that for all $\epsilon > 0$ there is a positive integer n_0^k such that for all $n \geq n_0^k$: $|F_n^k(x) - f(x)| < \frac{\beta-\alpha}{D}\epsilon$ for all x. The finite constant $D > 0$ is the bound in (22) on the difference of the mixed likelihoods $|n_\alpha(y) - n_\beta(y)| < D$ for all y. Define the positive integer n_0 as $n_0 = \max\{n_0^1, \ldots, n_0^q\}$. The theorem follows from this

bound and Lemma 2 and the definition of the 2-bell-curve mixtures (12) - (13) because then for all $n \geq n_0$:

$$|q_n(y|x) - p(y|x)| = |\sum_{k=1}^{q} \lambda^k(x) q_n^k(y|x) - p(y|x)| \quad (23)$$

$$= |\sum_{k=1}^{q} \lambda^k(x) q_n^k(y|x) - \sum_{k=1}^{q} \lambda^k(x) p(y|x)| \quad (24)$$

$$\leq \sum_{k=1}^{q} \lambda^k(x) |q_n^k(y|x) - p(y|x)| \quad (25)$$

$$= \sum_{k=1}^{q} \lambda^k(x) |[v_n^k(x) n_\alpha(y) + (1 - v_n^k(x)) n_\beta(y)]$$
$$- [w(x) n_\alpha(y) + (1 - w(x)) n_\beta(y)]| \quad (26)$$

$$= \sum_{k=1}^{q} \frac{\lambda^k(x)}{\beta - \alpha} |(F_n^k(x) - f(x)) n_\beta(y)$$
$$- (F_n^k(x) - f(x)) n_\alpha(y)| \quad (27)$$

$$= \sum_{k=1}^{q} \frac{\lambda^k(x)}{\beta - \alpha} |F_n^k(x) - f(x)||n_\beta(y) - n_\alpha(y)| \quad (28)$$

$$< \sum_{k=1}^{q} \frac{\lambda^k(x)}{\beta - \alpha} |F_n^k(x) - f(x)| D \quad (29)$$

$$< \sum_{k=1}^{q} \frac{\lambda^k(x)}{\beta - \alpha} \frac{\beta - \alpha}{D} \epsilon D \quad (30)$$

$$= \epsilon \sum_{k=1}^{q} \lambda^k(x) \quad (31)$$

$$= \epsilon \quad (32)$$

because $\alpha = \min\{\alpha^1, \ldots, \alpha^q\} < \beta = \max\{\beta^1, \ldots, \beta^q\}$ and because the coefficients $\lambda^k(x) \geq 0$ are convex and so $\sum_{k=1}^{q} \lambda^k(x) = 1$ for all x. So for all $n \geq n_0 = \max\{n_0^1, \ldots, n_0^q\}$: $|q_n(y|x) - p(y|x)| < \epsilon$ holds *both* for all x and for all y. So $q_n(y|x)$ converges uniformly to $p(y|x)$. **Q.E.D.**

The boundedness of the combined fuzzy systems F_n^k extends Theorem 1 to the far richer case of uniform convergence of any continuously transformed system $\phi^k(F_n^k)$ for *any* continuous real function ϕ^k on the closed and bounded and thus compact domain $[-B, B]$ because then ϕ^k is uniformly continuous [31]. So learning a given system F_n^k lets the user thereby learn any $\phi^k(F_n^k)$ as well [32]. This result also applies to throughput combination of fuzzy subsystems F_n^k as we now sketch. A result in analysis [33] further lets us take a continuous function of any finite number of continuous functions and still get a continuous function. So composition of these continuous functions with the uniformly converging bounded fuzzy systems still gives uniform mixture convergence of the continuously transformed fuzzy systems.

IV. Uniform Convergence of Mixtures of *THROUGHPUT*-Combined Fuzzy Systems

Uniform mixture convergence still holds for the throughput combination of any finite number q of additive fuzzy systems $F^1, \ldots, F^q$. We first review how the additive structure of the fuzzy systems $F_n^1, \ldots, F_n^q$ leads to a simple convex structure for the combined system F_n and its moments and its Bayesian posteriors over its subsystems and over their rules. This convex structure further lets us use Theorem 1 to prove the uniform mixture convergence in the richer case of throughput combination in Theorem 2 below.

Throughput combination combines the weighted rule firings $v^k b^k(y|x)$ of the q additive fuzzy systems F^k and then computes an output given a vector input x. Here the nonnegative weights $v^1, \ldots, v^q$ weight the respective fuzzy subsystems $F^1, \ldots, F^q$ and may depend on x and on other parameters. This gives the master rule firing $b(y|x)$ of the throughput-combined system F as

$$b(y|x) = \sum_{k=1}^{q} v^k b^k(y|x) = \sum_{k=1}^{q} \sum_{j=1}^{m_k} v^k b_j^k(y|x) \quad (33)$$

$$= \sum_{k=1}^{q} \sum_{j=1}^{m_k} v^k w_j^k a_j^k(x) b_j^k(y) \quad (34)$$

for standard additive fuzzy systems where input x fires the jth rule $R_{A_j^k \to B_j^k}$ of the kth subsystem F_k to degree $b_j^k(y|x) = w_j^k a_j^k(x) b_j^k(y)$ for rule weight w_j^k. Then the master mixture $p(y|x)$ of F has the convex form

$$p(y|x) = \frac{b(y|x)}{\int b(y|x) dy} = \frac{\sum_{k=1}^{q} \sum_{j=1}^{m_k} v^k b_j^k(y|x)}{\sum_{k=1}^{q} \sum_{j=1}^{m_k} v^k \int b_j^k(y|x)} \quad (35)$$

$$= \sum_{k=1}^{q} \sum_{j=1}^{m_k} p_j^k(x) p_{B_j^k}(y) \quad (36)$$

where the local prior or generalized mixture weight $p_j^k(x)$ is

$$p_j^k(x) = \frac{v^k a_j^k(x)\ w_j^k\ V_j^k}{\sum_{k=1}^{q} \sum_{l=1}^{m_k} v^k a_l^k(x)\ w_l^k\ V_l^k} \,. \quad (37)$$

Then the Bayesian posterior $p(j, k|y, x)$ over the m_k rules $R_{A_j^k \to B_j^k}$ of the kth combined fuzzy subsystem F^k is

$$p(j, k|x, y) = \frac{p_j^k(x) p_{B_j^k}(y)}{p(y|x)} \quad (38)$$

$$= \frac{p_j^k(x) p_{B_j^k}(y)}{\sum_{k=1}^{q} \sum_{j=1}^{m_k} p_j^k(x) p_{B_j^k}(y)} \,. \quad (39)$$

Marginalizing out the rule random variable gives the higher-level subsystem posterior $p(k|x, y)$ over the q subsystems $F^1, \ldots, F^q$ for the input x that produces output $y = F(x)$:

$$p(k|x, y) = \sum_{j=1}^{m_k} p(k, j|x, y) \quad (40)$$

$$= \frac{\sum_{j=1}^{m_k} p_j^k(x)\ p_{B_j}^k(y|x)}{\sum_{k=1}^{q} \sum_{j=1}^{m_k} p_j^k(x)\ p_{B_j}^k(y|x)} \,. \quad (41)$$

Figure 2 displays both the subsystem posterior $p(k|x, y)$ and the rule or telescoped posterior $p(j, k|x, y)$ for the 3 combined additive fuzzy systems in Figure 1. Further combining additively combined systems will give a new triple-summed master

mixture $p(y|x)$ with triple-summed rule posterior $p(j,k,l|x,y)$ and so on for any finite number of hierarchically combined additive fuzzy systems.

The throughput-combined fuzzy system F is just the first moment of the master mixture $p(y|x)$:

$$F(x) = E_{p(y|x)}[Y|X=x] = \int yp(y|x)dy \tag{42}$$

$$= \sum_{k=1}^{q}\sum_{j=1}^{m_k} p_j^k(x)c_j^k(x). \tag{43}$$

This throughput-combined fuzzy system F also arises through the ad hoc technique of identifying the centroid of the total rule firings $b(y|x)$ with the output $F(x)$: $F(x) = \text{Centroid}(B(y|x)) = \text{Centroid}(v^1B^1(y|x) + \cdots + v^qB^q(y|x))$. A telescoped second central moment or conditional variance arises over the subsystems and here over the total set of rules:

$$V[Y|X=x] = \sum_{k=1}^{q}\sum_{j=1}^{m_k} p_j^k(x)\sigma_{B_j(x)}^{2,k} + \sum_{k=1}^{q}\sum_{j=1}^{m_k} p_j^k(x)[c_j^k(x) - F(x)]^2 \ . \tag{44}$$

Theorem 2 shows that the uniform mixture approximation of Theorem 1 still holds for the case of throughput combination. The key idea of the proof is to reduce the throughput technique to a new form of convex combination of the combined systems $F^1, \ldots, F^q$: $F(x) = \sum_{k=1}^{q}\theta^k(x)F^k(x)$. Now the mixture sequence $q_n(y|x)$ in (14) has Watkins mixing coefficients $v_n(x)$ and $1 - v_n(x)$:

$$v_n(x) = \frac{\beta - F(x)}{\beta - \alpha} \ \text{ and } \ 1 - v_n(x) = \frac{F(x) - \alpha}{\beta - \alpha} \tag{45}$$

for the throughput combined fuzzy system F in (43).

Theorem 2: Uniform Mixture Convergence of *Throughput* Combined Bounded Fuzzy Systems.

The 2-bell-curve mixture sequence $q_n(y|x)$ in (45) of the throughput combined fuzzy system F_n in (43) converges uniformly to the 2-bell-curve mixture $p(y|x)$ in (11) of the target function f if each bounded nonconstant fuzzy system F_n^k converges uniformly to f.

Proof : Suppose that each bounded nonconstant fuzzy system F_n^k converges uniformly to f for each k. The result follows from Theorem 1 if we can write the throughput combined fuzzy system F_n as some convex combination $F(x) = \sum_{k=1}^{q}\theta^k(x)F^k(x)$ of the q fuzzy subsystems F_n^k.

The proof trick is a new normalization. Rewrite the master mixture $p(y|x)$ in (35) of the throughput-combined fuzzy system F (and thus of F_n for a sequence index n) as the equivalent subsystem-level convex sum $p(y|x) = \sum_{k=1}^{q}\theta^k(x)p^k(y|x)$:

$$p(y|x) = \frac{\sum_{k=1}^{q} v^k b^k(y|x)}{\sum_{k=1}^{q}\int v^k b^k(y|x)dy} \tag{46}$$

$$= \frac{\sum_{k=1}^{q}\int v^k b^k(y|x)dy\left[\frac{v^k b^k(y|x)}{\int v^k b^k(y|x)dy}\right]}{\sum_{k=1}^{q}\int v^k b^k(y|x)dy} \tag{47}$$

$$= \frac{\sum_{k=1}^{q}\int v^k b^k(y|x)dy}{\sum_{l=1}^{q}\int v^l b^l(y|x)dy}\ p^k(y|x) \tag{48}$$

$$= \sum_{k=1}^{q}\theta^k(x)p^k(y|x)\ . \tag{49}$$

Now the priors or convex coefficients $\theta^k(x)$ are

$$\theta^k(x) = \frac{\int v^k b^k(y|x)dy}{\sum_{l=1}^{q}\int v^l b^l(y|x)dy}. \tag{50}$$

So the generalized subsystem likelihoods $p^k(y|x)$ are

$$p^k(y|x) = \frac{v^k b^k(y|x)}{\int v^k b^k(y|x)dy} = \frac{b^k(y|x)}{\int b^k(y|x)dy} \tag{51}$$

for any subsystem weight $v^k > 0$ since v^k does not depend on y. The kth subsystem F^k is just the conditional mean $E_{p^k(y|x)}[Y|X=x]$ with respect to the likelihood $p^k(y|x)$. Then taking expectations with respect to $p(y|x)$ at iteration n gives $F_n(x) = \sum_{k=1}^{q}\theta^k(x)\ F_n^k(x)$. The result follows from Theorem 1 by putting $\theta^k(x) = \lambda^k(x)$ for all k and all x. So $q_n(y|x)$ converges uniformly to $p(y|x)$. $\mathbf{Q.E.D.}$

The new convex coefficient $\theta^k(x)$ in (50) sums all the kth subsystem's priors $p_j^k(x)$ in F^k:

$$\theta^k(x) = \frac{v^k\int b^k(y|x)dy}{\sum_{l=1}^{q} v^l\int b_j^l(y)dy} \tag{52}$$

$$= \frac{v^k\sum_{j=1}^{m_k} a_j^k(x)w_j^k\int b_j^k(y)dy}{\sum_{l=1}^{q} v^l\sum_{i=1}^{m_l} a_i^l(x)w_i^l\int b_i^l(y)dy} \tag{53}$$

$$= \sum_{j=1}^{m_k}\frac{v^k a_j^k(x)w_j^k V_j^k}{\sum_{l=1}^{q}\sum_{i=1}^{m_l} v^l a_i^l(x)w_i^l V_i^l} \tag{54}$$

$$= \sum_{j=1}^{m_k} p_j^k(x) \tag{55}$$

from (37). So the new convex weight $\theta^k(x)$ contains the complete rule-throughput information of $F^1, \ldots, F^q$ after all even though it simply weights the kth lone subsystem mixture $p^k(y|x)$ in the master mixture $p(y|x)$: $F(x) = \int yp(y|x)dy = \sum_{k=1}^{1}\theta^k(x)\int yp^k(y|x)dy = \sum_{k=1}^{q}\theta^k(x)F^k(x)$.

The convex weights θ^k in (54) also show that throughput combination defines q *meta-rules* $\mathcal{R}_{X\to Y}^{F^1}, \ldots, \mathcal{R}_{X\to Y}^{F^q}$. The output $F(x)$ combines the q total rule firings per subsystem through the q convex mixing weights $\theta^k(x) = \frac{\mathcal{A}^k(x)}{\sum_{i=1}^{q}\mathcal{A}^i(x)}$ if $\mathcal{A}^k = \sum_{j=1}^{m_k} v^k a_j^k(x)w_j^k V_j^k$. So $\theta^k(x)$ describes how x fires the kth meta-rule $\mathcal{R}_{X\to Y}^{F^k}$ in direct ratio analogy to how the lone mixing weight $p_j^k(x)$ describes how x fires the jth rule $R_{A_j^k\to B_j^k}$ in F^k.

V. Conclusions

Combining q fuzzy systems $F^1, \dots, F^q$ can often improve how well the combined rule-based system F approximates a sampled target function f. The target function f can be a trained deep neural classifier or any other source of representative input-output training data for the adaptive fuzzy systems.

The m_k if-then rules $R_{A_j^k \to B_j^k}$ of a given fuzzy system F^k define a generalized probability mixture $p^k(y|x)$ that gives back the system F^k as its first moment and whose higher moments describe the inherent uncertainty of the rule-based system.

The rules of any number q of fuzzy systems F^k define a higher-order probability mixture $p(y|x)$ that describes both the q individual subsystems F^k and the performance of each of their rules. The combination technique scales because ultimately mixing mixtures produces a new mixture. The older min-max fuzzy systems do not produce a probability mixture because they lack the needed additive structure.

Uniform convergence of the combined systems gives uniform convergence of both their direct mixtures and the mixture that describes the combined system. It also gives uniform convergence of the combined system's two-normal-curve mixtures. This gives in turn the uniform convergence of the mixtures that describe their continuous transformations. That holds because the boundedness of each combined fuzzy system gives uniform boundedness of the sequences of fuzzy systems and that gives uniform continuity of the transformations.

References

[1] Bart Kosko. Fuzzy Knowledge Combination. *International Journal of Intelligent Systems*, I:293–320, 1986.

[2] Ronald R. Yager. On ordered weighted averging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18(1):183–190, 1988.

[3] David L Hall and James Llinas. An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23, 1997.

[4] Josef Kittler, Mohamad Hatef, Robert PW Duin, and Jiri Matas. On combining classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 20(3):226–239, 1998.

[5] Didier Dubois, Henri Prade, and Ronald Yager. Merging fuzzy information. In *Fuzzy sets in approximate reasoning and information systems*, pages 335–401. Springer, 1999.

[6] Christian Wagner, Timothy C Havens, and Derek T Anderson. The arithmetic recursive average as an instance of the recursive weighted power mean. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2017.

[7] Yuan Li, Benjamin Rubinstein, and Trevor Cohn. Exploiting worker correlation for label aggregation in crowdsourcing. In *International Conference on Machine Learning*, pages 3886–3895. PMLR, 2019.

[8] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

[10] Mário Popolin Neto and Fernando V Paulovich. Explainable matrix-visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1427–1437, 2020.

[11] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021.

[12] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

[13] Wojciech Samek. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

[14] Ilaria Tiddi and Stefan Schlobach. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627, 2022.

[15] Akash Kumar Panda and Bart Kosko. Bayesian pruned random rule foams for XAI. In *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2021.

[16] Cátia M Salgado, Carlos S Azevedo, Jonathan Garibaldi, and Susana M Vieira. Ensemble fuzzy classifiers design using weighted aggregation criteria. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–5. IEEE, 2015.

[17] Lotfi A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[18] Richard Bellman, Robert Kalaba, and Lotfi Zadeh. Abstraction and pattern classification. *Journal of Mathematical Analysis and Applications*, 13:1–7, 1966.

[19] Tomohiro Takagi and Michio Sugeno. Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics*, (1):116–132, 1985.

[20] Gang Feng. A survey on analysis and design of model-based fuzzy control systems. *IEEE Transactions on Fuzzy systems*, 14(5):676–697, 2006.

[21] Anh-Tu Nguyen, Tadanari Taniguchi, Luka Eciolaza, Victor Campos, Reinaldo Palhares, and Michio Sugeno. Fuzzy control systems: Past, present and future. *IEEE Computational Intelligence Magazine*, 14(1):56–68, 2019.

[22] Toshiro Terano, Kiyoji Asai, and Michio Sugeno. *Fuzzy systems theory and its applications*. Academic Press Professional, Inc., 1992.

[23] Bart Kosko. Additive Fuzzy Systems: From Generalized Mixtures to Rule Continua. *International Journal of Intelligent Systems*, 33(8):1573–1623, 2018.

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

[25] Vladik Kreinovich, George C Mouzouris, and Hung T Nguyen. Fuzzy rule based modeling as a universal approximation tool. In *Fuzzy Systems*, pages 135–195. Springer, 1998.

[26] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[27] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[28] M. Jordan and T. Mitchel. Machine learning: Trends ,Perspectives, and Prospects. *Science, vol. 349*, pages 255–260, 2015.

[29] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer Science & Business Media, 2013.

[30] Fred Watkins. The representation problem for additive fuzzy systems. In *Proceedings of the International Conference on Fuzzy Systems (IEEE FUZZ-95)*, pages 117–122, 1995.

[31] James Munkres. Topology, 2014.

[32] Bart Kosko. Uniform mixture convergence of continuously transformed fuzzy systems. In *North American Fuzzy Information Processing Society Annual Conference*, pages 203–216. Springer, 2021.

[33] Walter Rudin. *Real and complex analysis*. McGraw-hill education, 2006.