Uniform Mixture Convergence of Continuously Transformed Fuzzy Systems

Bart Kosko(✉)

Department of Electrical and Computer Engineering,
University of Southern California, Los Angeles, CA, USA
kosko@usc.edu

Abstract. The probability mixture structure of additive fuzzy systems allows uniform convergence of the generalized probability mixtures that represent the if-then rules of one system or of many combined systems. A new theorem extends this result and shows that it still holds uniformly for any continuous function of such fuzzy systems if the underlying functions are bounded. This allows fuzzy rule-based systems to approximate a far wider range of nonlinear behaviors for a given set of sample data and still produce an explainable probability mixture that governs the rule-based proxy system.

1 Rule-Based Probability Mixtures and XAI

A new convergence theorem extends the scope of probabilistic mixture descriptions of fuzzy and other function approximators.

The new theorem shows that the uniform convergence of fuzzy rule-based approximators F_n carries over to their representing probability mixtures $q_n(y|x)$ for any continuous function ϕ of the fuzzy systems. So $F_n \to f$ uniformly for some bounded target function f implies not only that $\phi(F_n) \to \phi(f)$ uniformly. It also implies that $q_n(y|x) \to p_\phi(y|x)$ uniformly where now the Gaussian probability mixture $q_n(y|x)$ exactly represents $\phi(F_n)$ on average and where the Gaussian mixture $p_\phi(y|x)$ exactly represents $\phi(f)$ on average. Both Gaussian mixtures mix just *two* normal bell curves as in Fig. 1.

The mixture convergence theorem allows the same trained fuzzy system to model a much wider range of functions from the same training data while still using the explainability structure of the governing probability mixtures. These continuous functions can include norms and functions of norms and many other functions of the outputs of neural networks or other black-box approximators.

The converging mixtures track the convergence of the underlying fuzzy rule-based systems. Each additive fuzzy system F_n sums and averages its m fired if-then rules $R_{A_1 \to B_1}, \ldots, R_{A_m \to B_m}$ for each vector input x. The jth rule associates the then-part fuzzy set B_j with the if-part fuzzy set A_j. The fuzzy system's corresponding governing probability mixture $p_n(y|x)$ mixes m rule likelihood probabilities $p_{B_j}(y|x)$ with m convex mixing weights or prior densities

J. Rayz et al. (Eds.): NAFIPS 2021, LNNS 258, pp. 203–216, 2022.
https://doi.org/10.1007/978-3-030-82099-2_19

$p_j(x)$: $p_n(y|x) = p_1(x)p_{B_1}(y|x) + \cdots + p_m(x)p_{B_m}(y|x)$ [12]. The fuzzy systems F_n can converge to some sampled neural classifier or to any other black-box approximator.

The structured fuzzy system F acts as a *proxy* system for the otherwise inscrutable neural black box. The fuzzy proxy system's m rules $R_{A_j \to B_j}$ are inherently modular and their mixture structure further gives a statistical explanation of their operation. The proxy system can also combine q-many fuzzy systems $F^1, \ldots, F^q$ and each of these subsystems has its own mixture and its own rules.

The mixture structure endows both the fuzzy proxy system and the underlying sampled black box with a form of XAI or explainable AI [1,22,24,25]. This probabilistic description gets more accurate as the fuzzy system F_n at iteration n converges to the sampled neural network N. The probability description includes a complete Bayesian posterior probability $p(j|y,x)$ over the rules for each input x that fires the system. It also includes the higher-order moments such as the conditional variance that describes the system's uncertainty based on what the system has learned and based on which of the if-then rules the current input x fired.

The uniform convergence of additive fuzzy systems [9,10,15] lets the sequence of fuzzy systems F_n converge to or near a sufficiently sampled neural network. Uniform convergence lets the user pick an error tolerance level ϵ in advance that holds for all inputs x. Feedforward multilayer neural classifiers are bounded. Both such classifiers and neural regressors can also uniformly approximate continuous functions on compact sets if their hidden units are sigmoidal [2,8] or in some cases even if they are quasi-linear [4].

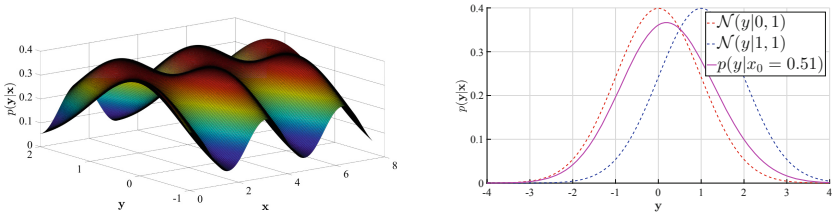


Fig. 1. Gaussian 2-bell-curve mixture representation for the continuously transformed target function $f(x) = \sin(x)$: $\phi(f(x)) = \sin^2(x)$. The generalized Gaussian mixture is $p_\phi(y|x) = w_\phi(x)N_{\alpha_\phi}(y|\alpha_\phi, \sigma_\alpha^2) + (1 - w_\phi(x))N_{\beta_\phi}(y|\beta_\phi, \sigma_\beta^2)$ from (9) where $\alpha_\phi = \inf_{x \in X} \phi(f(x)) = 0$ and $\beta_\phi = \sup_{x \in X} \phi(f(x)) = 1$. The first panel shows the mixture surface whose average is $\sin^2(x)$: $E_{p_\phi}[Y|X = x] = \sin^2(x)$. The second panel shows the two bell-curve likelihoods centered at α_ϕ and β_ϕ. The purple curve shows the particular Gaussian mixture $p(y|.51)$ that results if the input is $x = .51$.

Figure 1 shows the mixture surface and the two mixed Gaussian bell curves that exactly represent the continuously transformed target function $\phi(f(x)) =$

$f^2(x) = \sin^2(x)$ on average. The figure reflects the representation result in Theorem 1. Mixing two normal bell curves gives a generalized probability mixture $p_\phi(y|x)$ such that $\phi(f(x))$ equals the conditional mean $E_{p_\phi}[Y|X = x]$ with respect to $p_\phi(y|x)$ if the function f is bounded and ϕ is continuous: $E_{p_\phi}[Y|X = x] = \sin^2(x)$. The two mixed likelihood bell curves correspond roughly to the two then-part sets of a two-rule additive fuzzy system. This holds exactly when the bell-curve variances converge to zero and thus the normal densities converge to Dirac delta pulses centered at the set centroids. This limiting case reflects the older practice of picking then-part sets as spike centered at centroids.

Figures 2 and 3 illustrate the convergence result in Theorem 2. Figure 3 shows the mixture surfaces that represent 3 different transformed 20-rule fuzzy systems. Each fuzzy system approximates the transformed target function $f^2(x) = \sin^2(x)$ after the fuzzy systems have converged. The 20 rules $R_{A_1 \to B_1}, \ldots, R_{A_{20} \to B_{20}}$ in each system correspond to the 20 mixed likelihood probabilities p_{B_j} in each system's controlling mixture $p(y|x) = p_1(x)p_{B_1}(y) + \cdots + p_{20}(x)p_{B_{20}}(y)$. Figure 2 shows what the 2-bell-curve mixture $q_n(y|x)$ of each converging fuzzy system F_n in Fig. 2 looks like after $n = 4,000$ epochs of supervised learning. The 3 types of if-part fuzzy sets require 3 different supervised learning laws [11,16,19]. Figure 4 shows the Bayesian rule-posterior histograms for 4 different inputs to the 20-rule Gaussian fuzzy system in Fig. 3.

Figure 5 shows how randomly sampling from the Gaussian fuzzy system's trained 2-bell-curve mixture $p(y|x)$ can reproduce the transformed target function $\sin^2(x)$ through Monte Carlo averaging. This amounts to drawing a finite number of new if-then rules for each input x from a virtual *rule continuum* [12]. The rule histograms in Fig. 4 show that only a few of the stored or virtual rules fire for a given input x. This helps reduce the sampling costs of Monte Carlo averaging in high dimensions.

The next section presents the basic mathematical facts of the mixture approach to rule-based systems.

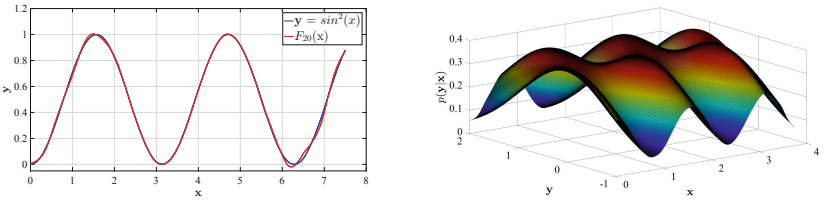
2 Probability Structure of Additive Fuzzy Rule-Based Systems

The new mixture convergence theorem exploits the fact that a generalized probability mixture $p(y|x)$ of just two Gaussian bell curves can exactly represent any bounded real function f as the average of the mixture: $f(x) = E[Y|X = x]$ for all x [12,13]. The conditional expectation $E[Y|X = x]$ integrates or sums with respect to the conditional Gaussian mixture $p(y|x)$:

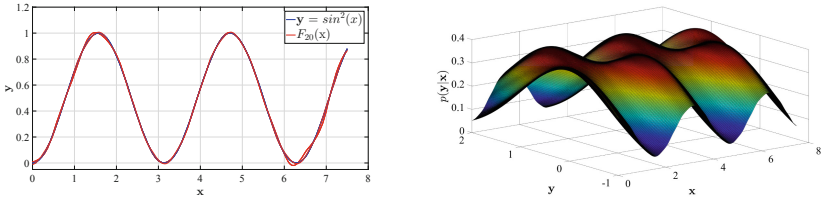
$$p(y|x) = w(x)N(y|\alpha, \sigma_\alpha^2) + (1 - w(x))N(y|\beta, \sigma_\beta^2) \tag{1}$$

for normal probability density $N(y|\alpha, \sigma_\alpha^2)$ with mean or location α and with any positive variance $\sigma_\alpha^2 > 0$ and likewise for $N(y|\beta, \sigma_\beta^2)$.

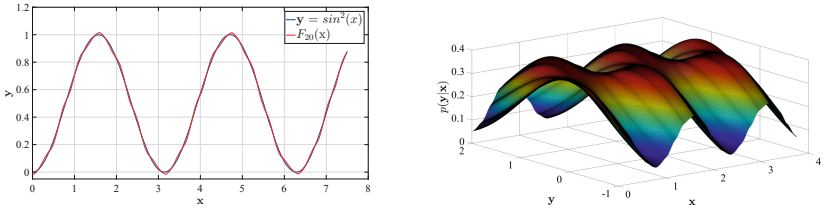
The normal densities are likelihoods. The convex mixing weights $w(x)$ and $1 - w(x)$ are priors or *Watkins coefficients* [11,26]. The mixing weights depend on



(a) 20 Sinc rules



(b) 20 Gaussian rules



(c) 20 Laplacian rules

Fig. 2. Two-bell-curve Gaussian mixture representation of 3 adaptive fuzzy systems after they have converged to the sampled continuously transformed bounded target function $\phi(f(x)) = \sin^2(x)$ Each fuzzy approximator used 20 rules but the mixture plots in panel (b) mixed just two Gaussian bell curves to produce the mixture density $q_n(y|x)$ whose average appears in panel (a). The plots show the mixtures and their averages after $n = 4,000$ epochs of supervised learning. The convergence plots illustrate the $q_n(y|x)$ convergence result in Theorem 2.

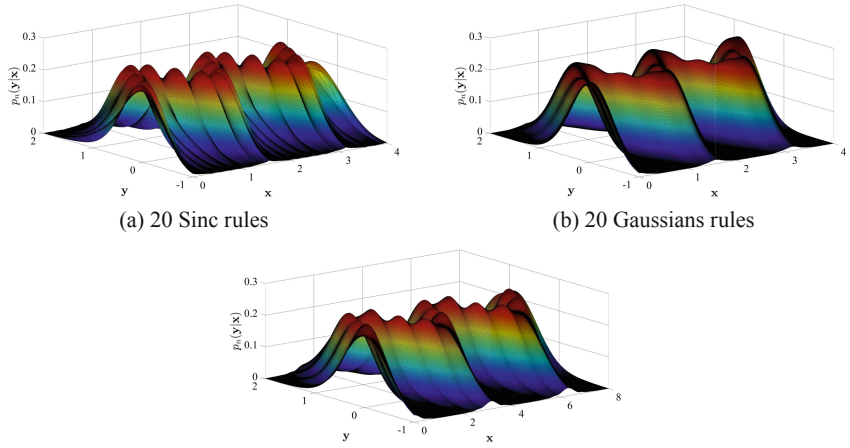
the input vector x through the bounded target function f with distinct bounds $\alpha \leq f \leq \beta$:

$$w(x) = \frac{\beta - f(x)}{\beta - \alpha}. \tag{2}$$

So the dual mixing weight is $1 - w(x) = \frac{f(x)-\alpha}{\beta-\alpha}$.

The mixture $p(y|x)$ is generalized because it depends on the input x. Ordinary mixtures $p(y)$ combine likelihoods that do not depend on x and that have constant convex mixing weights that also do not depend on x. The likelihoods can also depend on x but often do not in practice as we discuss below. The mixing weights always depend on x. The 2-bell-curve Gaussian mixture $p(y|x)$ in (1) gives an efficient way to represent a fuzzy or neural approximator and still have access to the mixture's XAI moment and Bayesian structure [13,20]. Figure 1

shows the mixture representation $p_\phi(y)|x)$ of the *square* of the target function $f(x) = \sin x$. The square-transformed function $\phi(f)$ composes the continuous square function ϕ with f to give $(\phi \circ f)(x) = \phi(f(x)) = \sin^2 x$.



(a) 20 Sinc rules

(b) 20 Gaussians rules

(c) 20 Laplacian rules.

Fig. 3. Mixture representations for the 3 adaptive fuzzy approximators $\phi(F_n)$ in Fig. 2 as they learned the continuously transformed target function $\phi(f(x)) = \sin^2(x)$. Each fuzzy approximator $F_n(x)$ used 20 rules and so their governing mixtures $p_n(y|x)$ each mixed 20 normal likelihoods with 20 convex mixing weights: $p_n(y|x) = p_1(x)p_{B_1}(y) + \cdots + p_{20}(x)p_{B_{20}}(y)$. The panels show each generalized mixture $p_n(y|x)$ after $n = 4,000$ epochs of learning.

This paper restricts the additive fuzzy systems to the important special case of SAMs or *standard additive model* fuzzy systems. Almost all fuzzy systems in practice are not just additive systems but are SAMs [5,11,18]. Older non-additive (min-max) fuzzy systems [3,23] do not give rise to a rule mixture.

SAMs scale rules to fire them. The input pattern vector x fires a SAM fuzzy system's if-then rule $R_{A_j \to B_j}$ by scaling the then-part fuzzy set B_j by the degree $a_j(x)$ to which x belongs to the corresponding if-part fuzzy set A_j. The jth rule $R_{A_j \to B_j}$ has the if-part fuzzy set $A_j \subset \mathbb{R}^n$ with membership or multivalued indicator function $a_j : \mathbb{R}^n \to [0, 1]$ so that $a_j(x) = \text{Degree}(x \in A_j)$. The rule has the corresponding then-part fuzzy set $B_j \subset \mathbb{R}$ with membership function $b_j : \mathbb{R} \to [0, 1]$. Then the jth fired then-part set $B_j(x)$ has SAM-scaled membership function $b_j(y|x) = a_j(x)b_j(y)$.

The additive system's total output set $B(y|x)$ sums and weights the m fired rule then-part sets $B_1(x), \ldots, B_m(x)$ for respective nonnegative rule weights $w_1, \ldots, w_m$. The weights w_j may depend on x or on other quantities in some applications. Then normalized rule firings define a generalized probability mixture [12] because we assume that the then-part set functions are nonnegative and integrable:

$$p(y|x) = \frac{b(y|x)}{\int b(y|x)dy} = \sum_{j=1}^{m} \frac{w_j a_j(x) V_j}{\sum_{k=1}^{m} w_k a_k(x) V_k} \frac{b_j(y)}{V_j} = \sum_{j=1}^{m} p_j(x) p_{B_j}(y) \quad (3)$$

with finite then-part set volume $V_j = \int b_j(y|x)dy > 0$ and convex mixing weights or priors $p_j(x) = \frac{w_j a_j(x) V_j}{\sum_{k=1}^{m} w_k a_k(x) V_k}$. The SAM rule-firing scaling $b_j(y|x) = a_j(x) b_j(y)$ gives the likelihood as $p_{B_j}(y|x) = p_{B_j}(y)$. Then the fuzzy system's output $F(x)$ is just the first noncentral moment of $p(y|x)$ and thus the output $F(x)$ is a convex combination of the rule then-part set centroids c_j:

$$F(x) = E[Y|X = x] = \int y\, p(y|x)\, dy = \sum_{j=1}^{m} p_j(x) c_j \quad (4)$$

where $c_j = \int y p_{B_j}(y)dy$. A conditional variance $V[Y|X = x]$ likewise describes the fuzzy system's second-order uncertainty for each input x.

Figure 3 shows the 3 20-rule mixture surfaces $p_n(y|x)$ that encode the squared target function $\phi(f(x)) = \sin^2 x$ for 3 different adaptive additive fuzzy systems based on 3 different types of rule if-part fuzzy sets: sinc, Gaussian, and Laplacian fuzzy sets. Random samples from the target function tune these parametrized sets along with other rule parameters [11]. The 2-bell-curve mixture compresses this system information in a multi-rule mixture into a simpler mixture $q_n(y|x)$ that can in turn mix with other such mixtures. Figure 2 shows this compression as the 2-bell-curve representations of the three converging fuzzy systems each approach the direct 2-bell-curve representation in Fig. 1.

The fuzzy system's governing mixture $p(y|x)$ itself states a form of the basic theorem on total probability. So it gives rise at once to a Bayesian posterior distribution $p(j|y, x)$ over the fuzzy system's m rules $R_{A_j \to B_j}$:

$$p(j|y, x) = p(R_{A_j \to B_j}|y, x) = \frac{p_j(x)\, p_{B_j}(y|x)}{p(y|x)} = \frac{p_j(x)\, p_{B_j}(y|x)}{\sum_{k=1}^{m} p_k(x)\, p_{B_k}(y|x)}. \quad (5)$$

The rule posterior $p(j|y, x)$ gives a complete description of the relative importance of each rule for each input x and observed output value $y = F(x)$. It is a powerful XAI tool that earlier rule-based systems simply ignored. Figure 4 shows the rule-posterior histograms for the 20-rule Gaussian fuzzy systems in Figs. 2 and 3 for 4 different inputs x after $n = 4,000$ epochs of supervised learning from the target function. This Bayesian posterior description becomes still more powerful when an adaptive fuzzy system approximates a neural black box because then the rule posterior gives a proxy posterior over the inner workings of the approximated black box. An additive fuzzy system can uniformly approximate any continuous (or bounded) target function on a compact domain [10,15]. This holds in practice if the fuzzy system trains with enough samples from the target function.

The posterior structure holds at a meta level if the fuzzy system F combines the rule *throughputs* of q fuzzy subsystems $F^1, \ldots, F^q$. Then the meta-level posterior $p(k|y, x)$ describes the relative firing of the q systems for each input x and

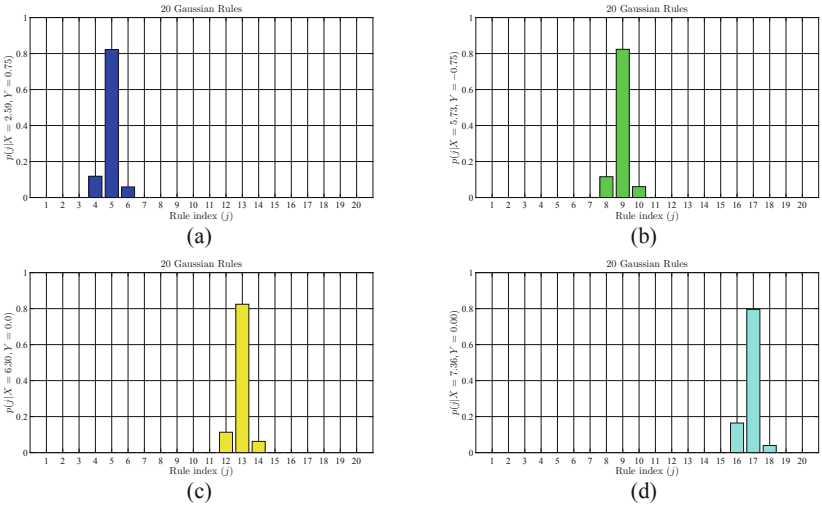


Fig. 4. Bayesian rule posteriors for the adaptive Gaussian fuzzy system in Figs. 2 and 3. The transformed fuzzy system $\phi(F_n)$ at iteration n used Gaussian if-part fuzzy sets for its 20 rules. The panels show the posterior $p(j|y,x)$ after $n = 4,000$ epochs of learning for 4 different input-output pairs (x,y). The posterior histograms show that just one rule contributed the most to the fuzzy system's rule-interpolated output $\phi(F_n(x))$. Most rules do not fire for a given input.

observed combined output y. The sub-level posterior $p(j,k|y,x)$ describes the relative rule firings of the m_k rules in the kth combined fuzzy system F^k:

$$p(j,k|y,x) = \frac{p_j^k(x)p_{B_j^k}(y)}{\sum_{k=1}^{q}\sum_{j=1}^{m_k} p_j^k(x)p_{B_j^k}(y)} \tag{6}$$

for the meta-level combined generalized mixture $p(y|x)$:

$$p(y|x) = \sum_{k=1}^{q}\sum_{j=1}^{m_k} p_j^k(x)p_{B_j^k}(y). \tag{7}$$

This telescoping posterior structure holds for any finite number of hierarchically combined additive fuzzy systems. Each layer adds another sum to the mixture.

The higher moments of a mixture $p(y|x)$ describe the higher-order statistical behavior of the rule-based proxy system. The conditional variance $V[Y|X=x]$ describes the uncertainty of a given output prediction $F(x)$ in terms of the inherent uncertainty in the then-parts of the if-then rules and the extent to which the output $F(x)$ interpolates over missing rules. The additive structure also gives a telescoping conditional variance when combining q additive fuzzy systems:

$$V[Y|X=x] = \sum_{k=1}^{q}\sum_{j=1}^{m_k} p_j^k(x)\sigma_{B_j^k}^2 + \sum_{k=1}^{q}\sum_{j=1}^{m_k} p_j^k(x)(c_j^k - F(x))^2 \tag{8}$$

where $\sigma_{B_j^k}^2$ is the variance of kth SAM's jth then-part fuzzy set B_j^k. These variance measures can help prune less-certain rules or help prune entire fuzzy subsystems as in random fuzzy foams [20].

A trained mixture $p(y|x)$ also lets one grow a fuzzy system by drawing if-then rules at random *from* the mixture. This implies sampling from a virtual rule *continuum* to estimate the output $F(x)$ for each input x. Figure 5 shows how such virtual rules can estimate the transformed target function $\sin^2 x$ by Monte Carlo averaging. The mixture becomes a compound in this case because the discrete rule index j becomes the continuous index θ [12]: $p(y|x) = \int_\Theta p_\theta(x) p_{B_\theta}(y) d\theta$.

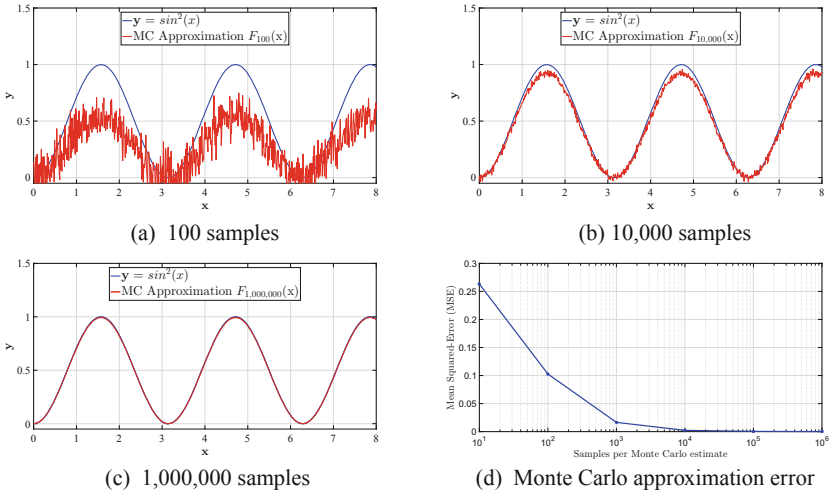
The continuous mixture exists so long as either density p_θ or p_{B_θ} is bounded. This sufficient condition holds because the bound $p_\theta(x) \leq d$ implies $p_\theta(x)p_{B_\theta}(y) \leq p_{B_\theta}(y)d$ and because that inequality integrates to $\int_\Theta p_\theta(x)p_{B_\theta}(y)d\theta \leq d$ for positive constant d since p_{B_θ} is a density. The fuzzy system's output $F(x)$ is again just the realized conditional expectation but with respect to the rule-continuous mixture $p(y|x)$: $F(x) = E[Y|X=x]$. So Monte Carlo approximates the output $F(x)$ as it uses the law of large numbers to approximate the expectation [7]. Each input x requires its own Monte Carlo sampling estimate to approximate the output $F(x)$.

Monte Carlo estimation does not depend on the input dimension but it does converge slowly [6]. The figure confirms that the standard error in the estimate falls off with the inverse square root of the number n of samples or rules drawn at random from $p(y|x)$ for a fixed x. So this virtual-rule technique carries a new computational burden to estimate $F(x)$. But it can mitigate rule explosion in high dimensions and it allows one to work with an estimated mixture $p(y|x)$.

3 Uniform Convergence of Mixtures of Transformed Fuzzy Systems

We now show first that a 2-bell-curve Gaussian mixture $p(y|x)$ can exactly represent any continuously transformed bounded real function f for any continuous function ϕ. This result extends the recent result of representing just a bounded target function f [13]. We present this and other results only for the scalar-valued case even though they extend componentwise to vector-valued systems.

Let $f : X \to \mathbb{R}$ be any bounded real function. So there exists a constant $B_f > 0$ such that $|f(x)| \leq B_f$ for all x. Let $\phi : \mathbb{R} \to \mathbb{R}$ be any continuous real function. The function ϕ is continuous at each real value x_0 just in case for all $\epsilon > 0$ there is a $\delta = \delta(\epsilon, x_0) > 0$ such that $|\phi(x_0) - \phi(x)| < \epsilon$ if $|x_0 - x| < \delta$ for all real x. Define the infimum $\alpha = \inf f$ and supremum $\beta = \sup f$ of the bounded function f. Assume that f is not constant and so $\alpha < \beta$ even though the results below still hold for constant functions if we subtract some constant $c > 0$ from α and add it to β. Assume likewise that $\alpha_\phi = \inf \phi(f) < \beta_\phi = \sup \phi(f)$ in all



(a) 100 samples (b) 10,000 samples

(c) 1,000,000 samples (d) Monte Carlo approximation error

Fig. 5. Monte Carlo fuzzy-system approximation of the continuously transformed target function $\phi(f(x)) = \sin^2(x)$ for the transformed rule-continuum fuzzy system $\phi(F)$ based on sampling from the mixture $p_\phi(y|x) = \frac{\beta - \phi(f(x))}{\beta - \alpha} \frac{1}{\sqrt{2\pi}} \exp[-\frac{(y-\alpha)^2}{2}] + \frac{\phi(f(x)) - \alpha}{\beta - \alpha} \frac{1}{\sqrt{2\pi}} \exp[-\frac{(y-\beta)^2}{2}]$ if $\alpha = \inf_{x \in X} \phi(f(x)) = 0$ and $\beta = \sup_{x \in X} f(x) = 1$ and unit variances. The blue lines in panels (a)-(c) show the transformed target $\phi(f(x)) = \sin^2(x)$. The red lines plot the Monte Carlo estimates of the transformed fuzzy system. Panel (a) plots the sample average of 100 y values drawn at random from $[0, 1]$ for each of 8,000 input values x 0.01 apart. Panel (b) plots the better approximation for 10,000 samples at each input x value. Panel (c) plots the still finer approximation for 1,000,000 such samples. Panel (d) shows the Monte Carlo approximation's slow inverse-square-root decay of the average squared error. Each plotted point averaged 10 runs.

the results that follow. The boundedness of f ensures that $\phi(f)$ is bounded on the restricted compact domain $[-B, B]$ since the continuous image of a compact set is itself compact [17] and hence bounded.

Theorem 1 shows how a 2-bell-curve Gaussian mixture $p_\phi(y|x)$ can exactly represent $\phi(f)$ on average for any bounded target function f. It technically assumes only that $\alpha_\phi < \beta_\phi$. It replaces the raw Watkins coefficients w and $1 - w$ in (2) with

$$w_\phi(x) = \frac{\beta_\phi - \phi(f(x))}{\beta_\phi - \alpha_\phi}. \tag{9}$$

with dual convex mixing weight $1 - w_\phi(x) = \frac{\phi(f(x)) - \alpha_\phi}{\beta_\phi - \alpha_\phi}$.

Theorem 1: Gaussian Mixture Representation of Continuously Transformed Bounded Functions.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ *is bounded and that* $\alpha_\phi = \inf_{x \in X} \phi(f) < \beta_\phi = \sup_{x \in X} \phi(f)$ *for some continuous function* $\phi : \mathbb{R} \to \mathbb{R}$. *Suppose that the generalized Gaussian mixture density* $p_\phi(y|x)$ *mixes two normal probability density functions with Watkins coefficients* w_ϕ and $1 - w_\phi$ *in (9):*

$$p_\phi(y|x) = w_\phi(x) N_{\alpha_\phi}(y|\alpha_\phi, \sigma_\alpha^2) + (1 - w_\phi(x)) N_{\beta_\phi}(y|\beta_\phi, \sigma_\beta^2) \; . \tag{10}$$

Then $p_\phi(y|x)$ *represents the composite continuous function* $\phi(f)$ *exactly on average:* $E_{p_\phi}[Y|X = x] = \phi(f(x))$ *for all* x.

Proof: The centroid of the normal random variable $Y \sim N_{\alpha_\phi}(y|\alpha_\phi, \sigma_\alpha^2) = n_{\alpha_\phi}(y)$ is its location parameter α_ϕ for an positive scale or variance $\sigma_\alpha^2 > 0$. The centroid of $n_{\beta_\phi}(y)$ is likewise β_ϕ. Then taking the conditional expectation with respect to the mixture density $p_\phi(y|x)$ gives the result:

$$E_\phi[Y|X = x] = \int y \, p_\phi(y|x) \, dy \tag{11}$$

$$= \left(\frac{\beta_\phi - \phi(f(x))}{\beta_\phi - \alpha_\phi}\right) \int y \, n_{\alpha_\phi}(y) \, dy + \left(\frac{\phi(f(x)) - \alpha_\phi}{\beta_\phi - \alpha_\phi}\right) \int y \, n_{\beta_\phi}(y) \, dy \tag{12}$$

$$= \left(\frac{\beta_\phi - \phi(f(x))}{\beta_\phi - \alpha_\phi}\right) \alpha_\phi + \left(\frac{\phi(f(x)) - \alpha_\phi}{\beta_\phi - \alpha_\phi}\right) \beta_\phi \tag{13}$$

$$= \frac{\phi(f(x))[\beta_\phi - \alpha_\phi]}{\beta_\phi - \alpha_\phi} \tag{14}$$

$$= \phi(f(x)) \tag{15}$$

since $\alpha_\phi < \beta_\phi$. **Q.E.D.**

We next state and prove two lemmas required to prove the main theorem: the 2-bell-curve Gaussian mixtures $q_n(y|x)$ that represent continuously transformed approximators $\phi(F_n)$ converge uniformly to the 2-bell-curve mixture $p_\phi(y|x)$ that represents the transformed target function $\phi(f)$. So uniform convergence of the transformed systems implies uniform convergence of their 2-bell-curve mixtures.

The two lemmas jointly show why we need only assume that the fuzzy approximators F_n are individually bounded if they converge uniformly to the target function f. This will imply that the continuously transformed approximators $\phi(F_n)$ converge uniformly to $\phi(f)$. This follows from the result of Lemma 1 that the uniform convergence of F_n to f promotes the individual boundedness of each F_n to the much stronger property of uniform boundedness.

Lemma 1: Bounded functions are uniformly bounded if they converge uniformly.

Suppose that each $F_n : \mathbb{R}^n \to \mathbb{R}$ *is bounded and that* F_n *converges uniformly to* f. *Then the functions* $\{F_n\}$ *are uniformly bounded and* f *is bounded.*

Proof: Say that F_n is bounded. So there is a bound $B_n > 0$ such that $|F_n(x)| \leq B_n$ for all x. Suppose that F_n converges uniformly to f. Then for all $\epsilon > 0$ there

is a positive integer n_0 such that for all $n \geq n_0$: $|F_n(x) - f(x)| < \epsilon$ for all x. Then we can bound the tail of the sequence with the triangle inequalities for all $n \geq n_0$ and for all x: $|F_n(x)| - |F_{n_0}(x)| \leq |\, |F_n(x)| - |F_{n_0}(x)| \,| \leq |F_n(x) - F_{n_0}(x)| \leq |F_n(x) - f(x)| + |f(x) - F_{n_0}(x)| < 2\epsilon$ from uniform convergence. Take $\epsilon = \frac{1}{2}$ for convenience since $\epsilon > 0$ was arbitrary. This gives the tail bound: $|F_n(x)| < 1 + |F_{n_0}(x)| \leq 1 + B_{n_0}$ since F_{n_0} is bounded. Put $B = \max(B_1, \ldots, B_{n_0 - 1}, 1 + B_{n_0})$. Then $|F_n(x)| \leq B$ for all n and all x. So $\{F_n\}$ is uniformly bounded.

The target function f inherits the boundedness of the functions F_n because of uniform convergence: $|f(x)| = |f(x) - F_n(x) + F_n(x)| \leq |F_n(x) - f(x)| + |F_n(x)| < \frac{1}{2} + B$ for all x from the uniform boundedness of the sequence $\{F_n\}$. **Q.E.D.**

Lemma 2 uses the crucial real-analytical fact that a continuous function on a compact set is *uniformly* continuous [17,21]. So then for all $\epsilon > 0$ there is a $\delta = \delta(\epsilon) > 0$ such that $|\phi(u) - \phi(v)| < \epsilon$ if $|u - v| < \delta$ for all u and all v. The uniform bound $B > 0$ in Lemma 1 for the functions $\{F_n\}$ gives the compact interval $[-B, B]$. So ϕ is uniformly continuous on $[-B, B]$. Then the δ condition lets the same ϵ hold for $|\phi(F_n(x)) - \phi(f(x))| < \epsilon$ for all x and thus gives uniform convergence of $\phi(F_n)$ to $\phi(f)$. The proof of Theorem 2 shows how this uniform convergence then leads to the uniform convergence of the 2-bell-curve mixtures $q_n(y|x)$ to $p_\phi(y|x)$.

Lemma 2: Uniform convergence of bounded functions implies uniform convergence of their continuous transformations.

Suppose that $\phi : \mathbb{R} \to \mathbb{R}$ is continuous. Suppose that each $F_n : \mathbb{R}^n \to \mathbb{R}$ is bounded and that F_n converges uniformly to f. Then $\phi(F_n)$ converges uniformly to $\phi(f)$.

Proof: Suppose that F_n is bounded and that F_n converges uniformly to f: For all $\epsilon > 0$ there is a positive integer n_0 such that for all $n \geq n_0$: $|F_n(x) - f(x)| < \epsilon$ for all x. Then Lemma 1 states that the sequence $\{F_n\}$ has the uniform bound $B > 0$: $|F_n(x)| \leq B$ for all n and for all x. So $-B \leq F_n(x) \leq B$ holds for all positive n. The real interval $[-B, B]$ is compact because it is closed and bounded. Then the restricted continuous function $\phi : [-B, B] \to \mathbb{R}$ on the compact interval $[-B, B]$ is uniformly continuous [17]. So for all $\epsilon > 0$ there is a $\delta > 0$ that depends only on ϵ and not any x and such that $|\phi(u) - \phi(v)| < \epsilon$ if $|u - v| < \delta$ for all u and for all v in $[-B, B]$. Put $u = F_n(x)$ and $v = f(x)$. Then $|\phi(F_n(x)) - \phi(f(x))| < \epsilon$ holds because $|F_n(x) - f(x)| = |u - v| < \delta$. So for every $\epsilon > 0$ there is a positive integer n_0 such that for all $n \geq n_0$: $|\phi(F_n(x)) - \phi(f(x))| < \epsilon$ for all x. So the composite function $\phi \circ F_n$ converges uniformly to the composite function $\phi \circ f$. **Q.E.D.**

Theorem 2 uses the 2-bell-curve Gaussian mixture $q_n(y|x)$ for each continuously transformed fuzzy approximator $\phi(F_n)$. Let $q_n(y|x)$ denote the 2-bell-curve Gaussian mixture for $\phi(F_n(x))$ for fuzzy system F_n with Watkins coefficients v_n and $1 - v_n$ of the transformed system $\phi(F_n)$:

$$v_n(x) = \frac{\beta_\phi - \phi(F_n(x))}{\beta_\phi - \alpha_\phi} \tag{16}$$

$$1 - v_n(x) = \frac{\phi(F_n(x)) - \beta_\phi}{\beta_\phi - \alpha_\phi} \tag{17}$$

such that $\alpha_\phi \leq \phi(F_n) \leq \beta_\phi$ for all n. Then the 2-bell-curve Gaussian mixture $q_n(y|x)$ has the form

$$q_n(y|x) = v_n(x)N_{\alpha_\phi}(y|\alpha_\phi, \sigma_\alpha^2) + (1 - v_n(x))N_{\beta_\phi}(y|\beta_\phi, \sigma_\beta^2) \; . \tag{18}$$

Theorem 2: Uniform Convergence of Gaussian Mixtures that Represent Continuously Transformed Systems.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ *is bounded and that* $\alpha_\phi = \inf_{x \in X} \phi(f) < \beta_\phi = \sup_{x \in X} \phi(f)$ *for some continuous function* $\phi : \mathbb{R} \to \mathbb{R}$ *and that* $\alpha_\phi \leq \phi(F_n(x)) \leq \beta_\phi$ *for all* n *and for all* x. *Suppose that the bounded additive fuzzy systems* F_n *uniformly converge to the target function* f. *Suppose that the 2-bell-curve Gaussian mixture density* $p_\phi(y|x)$ *that represents* $\phi(f)$ *mixes two normal probability density functions with Watkins coefficients* w_ϕ *and* $1 - w_\phi$ *in (9):*

$$p_\phi(y|x) = w_\phi(x)N_{\alpha_\phi}(y|\alpha_\phi, \sigma_\alpha^2) + (1 - w_\phi(x))N_{\beta_\phi}(y|\beta_\phi, \sigma_\beta^2) \; . \tag{19}$$

Suppose further that the Gaussian mixture $q_n(y|x)$ *with Watkins coefficients* v_n *and* $1 - v_n$ *in (16)–(17) that represents* $\phi(F_n)$ *has the like 2-bell-curve form*

$$q_n(y|x) = q_n(x)N_{\alpha_\phi}(y|\alpha_\phi, \sigma_\alpha^2) + (1 - v_n(x))N_{\beta_\phi}(y|\beta_\phi, \sigma_\beta^2) \; . \tag{20}$$

Then $q_n(y|x)$ *converges uniformly to* $p_\phi(y|x)$ *uniformly in* x *and* y.

Proof: The two normal bell-curve likelihoods $N_{\alpha_\phi}(y|\alpha_\phi, \sigma_\alpha^2)$ and $N_{\beta_\phi}(y|\beta_\phi, \sigma_\beta^2)$ are bounded functions for any fixed variances $\sigma_\alpha^2 > 0$ and $\sigma_\beta^2 > 0$. So $|n_{\alpha_\phi}(y) - n_{\beta_\phi}(y)| \leq \max(n_{\alpha_\phi}(\alpha_\phi), n_{\beta_\phi}(\beta_\phi))$ holds for all y values for the modes α_ϕ and β_ϕ. Then $|n_\alpha(y) - n_\beta(y)| < D$ if $D = \max(n_\alpha(\alpha), n_\beta(\beta)) + 1$.

Lemmas 1 and 2 and the boundedness of F_n imply that $\phi(F_n)$ converges uniformly to $\phi(f)$ because ϕ is continuous and because F_n converges uniformly to f. So for all $\epsilon > 0$ there is a positive integer n_0 such that for all integers $n \geq n_0$: $|\phi(F_n(x)) - \phi(f(x))| < \frac{\beta_\phi - \alpha_\phi}{D}\epsilon$ for all $x \in X$. Then using the Watkins coefficients w_ϕ and v_n for the two respective Gaussian mixtures $p_\phi(y|x)$ and $q_\phi(y|x)$ gives for $n \geq n_0$:

$$\begin{aligned}
|q_n(y|x) - p_\phi(y|x)| &= \frac{1}{\beta_\phi - \alpha_\phi}|(\beta_\phi - \phi(F_n(x)))n_{\alpha_\phi}(y) + (\phi(F_n(x)) - \alpha_\phi)n_{\beta_\phi}(y) \\
&\quad - (\beta_\phi - \phi(f(x)))n_{\alpha_\phi}(y) - (\phi(f(x)) - \alpha_\phi)n_{\beta_\phi}(y)| && (21) \\
&= \frac{1}{\beta_\phi - \alpha_\phi}|\phi(F_n(x))(n_{\beta_\phi}(y) - n_{\alpha_\phi}(y)) - \phi(f(x))(n_{\beta_\phi}(y) - n_{\alpha_\phi}(y))| && (22) \\
&= \frac{1}{\beta_\phi - \alpha_\phi}|\phi(F_n(x)) - \phi(f(x))||n_{\alpha_\phi}(y) - n_{\beta_\phi}(y)| && (23) \\
&< \frac{1}{\beta_\phi - \alpha_\phi}|\phi(F_n(x)) - \phi(f(x))| \; D && (24) \\
&< \frac{1}{\beta_\phi - \alpha_\phi}\frac{\beta_\phi - \alpha_\phi}{D}\epsilon \; D && (25) \\
&= \epsilon && (26)
\end{aligned}$$

for all x and all y. So $q_n(y|x)$ converges to $p_\phi(y|x)$ uniformly in x and y. **Q.E.D.**

Figure 2 illustrates Theorem 2 for 3 adaptive SAM approximators that each use 20 rules. The figure shows the 3 different adaptive-SAM 2-bell-curve Gaussian mixtures $q_n(y|x)$ after they have converged to the transformed target function's 2-bell-curve Gaussian mixture $p_\phi(y|x)$.

4 Conclusions

The additive structure of a fuzzy rule-based system F with m if-then rules $R_{A_1 \to B_1}, \ldots, R_{A_m \to B_m}$ yields a controlling generalized mixture $p(y|x)$ that mixes m likelihood probability densities: $p(y|x) = p_1(x)p_{B_1}(y|x) + \cdots + p_m(x)p_{B_m}(y|x)$. The mixture's first noncentral moment $E_p[Y|X]$ gives back the fuzzy system F as $F(x) = E_p[Y|X = x]$ for all inputs x. So the mixture structure itself avoids the many earlier *ad hoc* definitions of F and reveals a natural and useful connection to probability theory. The mixture's higher moments give a statistical description of the system's uncertainty for each input x based on what the system has learned in its encoded rule structure. The mixture's convex structure further gives a Bayesian posterior probability that describes the m rules and any of its rule-based subsystems.

These mixture properties allow an adaptive additive fuzzy system to learn a rule-based and statistically explainable proxy version of a sampled neural black box. A 2-bell-curve Gaussian mixture $p_\phi(y|x)$ can further represent any continuously transformed bounded scalar function $\phi(f)$. The result extends componentwise to vector-valued functions.

A new convergence theorem shows that the uniform convergence of continuously transformed fuzzy systems $\phi(F_n)$ to a bounded continuously transformed target function or black box $\phi(f)$ implies that the generalized 2-bell-curve Gaussian mixtures $q_n(y|x)$ that characterize the transformed fuzzy systems $\phi(F_n)$ converge uniformly to the like mixture $p_\phi(y|x)$ that characterizes the transformed target $\phi(f)$. A major research challenge is to extend these rule-based convergence results to the XAI proxy modeling of converging *feedback* neural dynamical systems as in [14].

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control Signals Syst. **2**(4), 303–314 (1989). https://doi.org/10.1007/BF02551274
3. Dubois, D., Hüllermeier, E., Prade, H.: A systematic approach to the assessment of fuzzy association rules. Data Min. Knowl. Discov. **13**(2), 167–192 (2006). https://doi.org/10.1007/s10618-005-0032-4
4. Elbrächter, D., Perekrestenko, D., Grohs, P., Bölcskei, H.: Deep neural network approximation theory. IEEE Trans. Inf. Theory **67**(5), 2581–2623 (2021)
5. Feng, G.: A survey on analysis and design of model-based fuzzy control systems. IEEE Trans. Fuzzy Syst. **14**(5), 676–697 (2006)

6. Glasserman, P.: Monte Carlo Methods in Financial Engineering, vol. 53. Springer, Heidelberg (2013). https://doi.org/10.1007/978-0-387-21617-1
7. Hogg, R.V., McKean, J., Craig, A.T.: Introduction to Mathematical Statistics. Pearson, London (2013)
8. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Netw. **2**(5), 359–366 (1989)
9. Kosko, B.: Neural Networks and Fuzzy Systems. Prentice-Hall, Hoboken (1991)
10. Kosko, B.: Fuzzy systems as universal approximators. IEEE Trans. Comput. **43**(11), 1329–1333 (1994)
11. Kosko, B.: Fuzzy Engineering. Prentice-Hall, Hoboken (1996)
12. Kosko, B.: Additive fuzzy systems: from generalized mixtures to rule continua. Int. J. Intell. Syst. **33**(8), 1573–1623 (2018)
13. Kosko, B.: Convergence of generalized probability mixtures that describe adaptive fuzzy rule-based systems. In: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8. IEEE (2020)
14. Kosko, B.: Bidirectional associative memories: unsupervised Hebbian learning to bidirectional backpropagation. IEEE Trans. Syst. Man Cybern. Syst. **51**(1), 103–115 (2021)
15. Kreinovich, V., Mouzouris, G.C., Nguyen, H.T.: Fuzzy rule based modeling as a universal approximation tool. In: Nguyen, H.T., Sugeno, M. (eds.) Fuzzy Systems. The Springer Handbook Series on Fuzzy Sets, vol. 2, pp. 135–195. Springer, Boston (1998). https://doi.org/10.1007/978-1-4615-5505-6_5
16. Mitaim, S., Kosko, B.: The shape of fuzzy sets in adaptive function approximation. IEEE Trans. Fuzzy Syst. **9**(4), 637–656 (2001)
17. Munkres, J.: Topology (2014)
18. Nguyen, A.-T., Taniguchi, T., Eciolaza, L., Campos, V., Palhares, R., Sugeno, M.: Fuzzy control systems: past, present and future. IEEE Comput. Intell. Mag. **14**(1), 56–68 (2019)
19. Osoba, O., Mitaim, S., Kosko, B.: Bayesian inference with adaptive fuzzy priors and likelihoods. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **41**(5), 1183–1197 (2011)
20. Panda, A.K., Kosko, B.: Random fuzzy-rule foams for explainable AI. In: Fuzzy Information Processing 2020, Proceedings of NAFIPS-2020. Springer, Heidelberg (2020). https://doi.org/10.1007/978-3-030-71098-9
21. Rudin, W.: Real and Complex Analysis. McGraw-Hill Education, New York (2006)
22. Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (eds.): Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNCS (LNAI), vol. 11700. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28954-6
23. Terano, T., Asai, K., Sugeno, M.: Fuzzy Systems Theory and its Applications. Academic Press Professional Inc., Cambridge (1992)
24. Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans. Neural Netw. Learn. Syst. pp. 1–21 (2020). https://doi.org/10.1109/TNNLS.2020.3027314
25. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating XAI: a comparison of rule-based and example-based explanations. Artif. Intell. **291**, 103404 (2021)
26. Watkins, F.: The representation problem for additive fuzzy systems. In: Proceedings of the International Conference on Fuzzy Systems (IEEE FUZZ-1995), pp. 117–122 (1995)