Uncertainty quantification in regression neural networks using likelihood-based belief functions

Thierry Denoeux

Uncertainty quantification in regression neural networks using likelihood-based belief functions

Thierry Denœux[0000−0002−0660−5436]

Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France
Institut universitaire de France, Paris, France
tdenoeux@utc.fr

Abstract. We introduce a new method for quantifying prediction uncertainty in regression neural networks using evidential likelihood-based inference. The method is based on the Gaussian approximation of the likelihood function and the linearization of the network output with respect to the weights. Prediction uncertainty is described by a random fuzzy set inducing a predictive belief function. Preliminary experiments suggest that the approximations are very accurate and that the method allows for conservative uncertainty-aware predictions.

Keywords: Evidence theory, Dempster-Shafer theory, machine learning, deep learning, random fuzzy sets.

1 Introduction

In recent years, research in machine learning (ML) has been increasingly focused on developing models that not only have good prediction performance, but also provide some measure of prediction uncertainty [1, 10]. The mainstream Bayesian approach is computationally intensive and it requires the existence of prior knowledge about the model parameters, an unrealistic assumption in the case of neural networks with thousands of weights. The Bayesian approach also does not clearly separate aleatory uncertainty (due to variability of the response given the predictors) from epistemic uncertainty (due to lack of knowledge of the true data distribution). In this paper, continuing previous work, I propose to explore another direction referred to as *evidential machine learning* (EML), in which uncertainty is quantified using belief functions. In particular, a belief function induced by a random set [13] or a random fuzzy set [7] has a probabilistic component, suitable for representing aleatory uncertainty, and a set-based component that can express epistemic uncertainty.

At least two main approaches have been proposed for supervised learning in the evidential ML framework. The *distance-based* approach consists in computing a predictive belief function by assessing the similarity between the input vector and training instances or prototypes. This idea was first proposed for classification [3][4][9] and was only recently applied to regression [5][6]; it does not assume any parametric statistical model. In contrast, the *likelihood-based* approach to statistical prediction, first introduced in [11] and revisited in [7]

using the new concept of epistemic random fuzzy set, starts with a parametric model and treats the relative likelihood function as a possibility distribution. By expressing the response variable as a function of the parameter and a random variable with known probability distribution, one obtains a random fuzzy set modeling prediction uncertainty. Noticeably, this approach boils down to Bayesian inference when a prior probability distribution is assumed.

Likelihood-based evidential prediction was applied to linear regression in [12] and to logistic regression in [8]. Applying it to nonlinear models with a large number of parameters while keeping computations tractable is particularly challenging. First results in this direction are reported in this paper, with a focus on regression neural networks. The rest of this paper is organized as follows. Necessary notions about random fuzzy sets and likelihood-based evidential inference will first be recalled in Section 2. Our new approach is then described in Section 3 and experimental results are reported in Section 4.

2 Background

Background notions about possibility theory and epistemic random fuzzy sets will first be recalled in Section 2.1. Evidential likelihood-based inference will then be summarized in Section 2.2.

2.1 Possibility Theory and Random Fuzzy Sets

Possibility and necessity measures. Let $\boldsymbol{\theta}$ be a variable taking values in Θ. Assume that we receive a piece of evidence telling us that "$\boldsymbol{\theta}$ is $\widetilde{F}$", where $\widetilde{F}$ is a normal fuzzy subset of Θ (i.e., a map $\widetilde{F}:\Theta\rightarrow[0,1]$ such that $\sup_{\theta\in\Theta}\widetilde{F}(\theta)=1$). This evidence induces a *possibility measure* $\Pi_{\widetilde{F}}$ from 2^{Θ} to $[0,1]$ defined by $\Pi_{\widetilde{F}}(B)=\sup_{\theta\in B}\widetilde{F}(\theta)$, for all $B\subseteq\Theta$. The number $\Pi_{\widetilde{F}}(B)$ is interpreted as the degree of possibility that $\boldsymbol{\theta}\in B$, given that $\boldsymbol{\theta}$ is $\widetilde{F}$ [16]. The corresponding *possibility distribution* is the mapping $\pi_{\widetilde{F}}:\Theta\rightarrow[0,1]$ defined by $\pi_{\widetilde{F}}(\theta)=\Pi_{\widetilde{F}}(\{\theta\})=\widetilde{F}(\theta)$. It is identical to $\widetilde{F}$: the degree of possibility that $\boldsymbol{\theta}=\theta$ given the flexible constraint "$\boldsymbol{\theta}$ is $\widetilde{F}$" is equal to the degree of membership of θ to fuzzy set $\widetilde{F}$. The dual *necessity measure* is defined as $N_{\widetilde{F}}(B)=1-\Pi_{\widetilde{F}}(B^c)$, where B^c denotes the complement of B in Θ.

Gaussian fuzzy vectors. A *Gaussian fuzzy vector (GFV)* is a normal fuzzy subset $\widetilde{F}$ of $\Theta=\mathbb{R}^p$ (with $p\geq 1$) such that $\widetilde{F}(\theta)=\exp\left(-\frac{1}{2}(\theta-m)^T\boldsymbol{H}(\theta-m)\right)$, where $m\in\mathbb{R}^p$ is the mode of $\widetilde{F}$, and $\boldsymbol{H}\in\mathbb{R}^{p\times p}$ is a symmetric and positive semidefinite precision matrix. We write $\widetilde{F}\sim\text{GFV}(m,\boldsymbol{H})$. When $p=1$, we say that $\widetilde{F}$ is a Gaussian fuzzy number (GFN). The following proposition (proved in [8]) states that the image of a GFV by a linear mapping is still a GFV.

Proposition 1 *Let $\boldsymbol{\theta}\in\mathbb{R}^p$ be a p-dimensional variable constrained by a possibility distribution $\pi_{\boldsymbol{\theta}}\sim GFV(m,\boldsymbol{H})$ with mode $m\in\mathbb{R}^p$ and positive definite precision matrix $\boldsymbol{H}\in\mathbb{R}^{p\times p}$. Let $\boldsymbol{U}\in\mathbb{R}^{q\times p}$ be a real matrix of rank*

$q \leq p$, $v \in \mathbb{R}^q$ *and* $\boldsymbol{z} = \boldsymbol{U}\theta + v \in \mathbb{R}^q$. *Variable* $\boldsymbol{z}$ *is constrained by* $\pi_{\boldsymbol{z}} \sim GFV\left(\boldsymbol{U}m + v, (\boldsymbol{U}\boldsymbol{H}^{-1}\boldsymbol{U}^T)^{-1}\right)$.

Random Fuzzy Sets. Let $(\Omega, \Sigma_\Omega, P)$ denote a probability space, (Θ, Σ_Θ) a measurable space, and $\widetilde{X}$ a mapping from Ω to the set $[0,1]^\Theta$ of fuzzy subsets of Θ. For any $\alpha \in [0,1]$, let ${}^\alpha\widetilde{X}$ be the mapping from Ω to 2^Θ such that $\omega \mapsto \{\theta \in \Theta : \widetilde{X}(\omega)(\theta) \geq \alpha\}$. If, for any $\alpha \in [0,1]$, ${}^\alpha\widetilde{X}$ is $\Sigma_\Omega - \Sigma_\Theta$ strongly measurable [13], the tuple $(\Omega, \Sigma_\Omega, P, \Theta, \Sigma_\Theta, \widetilde{X})$ is said to be a *random fuzzy set* (RFS) [2]. In Epistemic Random Fuzzy Set theory, a RFS represents a piece of evidence about a variable $\boldsymbol{\theta}$ taking values in Θ. The set Ω is seen as a *set of interpretations* of this piece of evidence, which may be unreliable, vague (fuzzy), or both. If interpretation $\omega \in \Omega$ holds, we only know that $\boldsymbol{\theta}$ is constrained by the possibility distribution defined by fuzzy set $\widetilde{X}(\omega)$. To any RFS verifying normalization conditions [7], we can associate a belief function representing one's beliefs based on the available evidence. For any $\omega \in \Omega$, a conditional possibility measure $\Pi_{\widetilde{X}(\omega)}$ on Θ can be defined as follows: for any $B \subseteq \Theta$, $\Pi_{\widetilde{X}(\omega)}(B) = \sup_{\theta \in B} \widetilde{X}(\omega)(\theta)$. For any $B \in \Sigma_\Theta$, let $Bel_{\widetilde{X}}(B)$ and $Pl_{\widetilde{X}}(B)$ denote, respectively, the *expected necessity* and the *expected possibility* of B wrt P. The corresponding mappings $Bel_{\widetilde{X}} : \Sigma_\Theta \to [0,1]$ and $Pl_{\widetilde{X}} : \Sigma_\Theta \to [0,1]$, are, respectively, belief and plausibility functions [2].

2.2 Evidential Likelihood-based Inference

We consider an observed random vector $\boldsymbol{Y}$ with probability density function (pdf) $f_{\boldsymbol{Y}|\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \Theta$ is the unknown parameter. The likelihood of any value θ of the parameter after observing $\boldsymbol{Y} = \boldsymbol{y}$ is $L(\theta; \boldsymbol{y}) = \eta f_{\boldsymbol{Y}|\theta}(\boldsymbol{y})$, where η is an arbitrary positive constant. Assuming that $\sup_\theta L(\theta; \boldsymbol{y}) < +\infty$, we can define the relative likelihood of θ as

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) = \frac{L(\theta; \boldsymbol{y})}{\sup_{\theta' \in \Theta} L(\theta'; \boldsymbol{y})}. \tag{1}$$

As proposed in [7], we interpret mapping $\pi_{\theta|\boldsymbol{y}} : \Theta \to [0,1]$ as a *possibility distribution* over Θ or, equivalently, as the *fuzzy set* of likely values of $\boldsymbol{\theta}$ after observing $\boldsymbol{Y} = \boldsymbol{y}$. It is, thus, a representation of the information about $\boldsymbol{\theta}$ provided by observation $\boldsymbol{y}$.

Assuming $\ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta)$ to be twice differentiable, a tractable approximation of function $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta)$ can often be obtained by computing a Taylor expansion of its logarithm about a solution $\widehat{\theta}$ of the score equation $\frac{\partial \ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}}{\partial \theta} = 0$ up to the second order [14]. We then obtain

$$\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) \approx \exp\left[-\frac{1}{2}(\theta - \widehat{\theta})^T \mathcal{I}(\widehat{\theta})(\theta - \widehat{\theta})\right], \tag{2}$$

where $\mathcal{I}(\widehat{\theta})$ is the *observed information matrix* defined as

$$\mathcal{I}(\widehat{\theta}) = -\left.\frac{\partial^2 \ln \pi_{\boldsymbol{\theta}|\boldsymbol{y}}}{\partial \theta \partial \theta^T}\right|_{\theta = \widehat{\theta}}.$$

As noted in [14], (2) is usually a good approximation when $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is an independent sample and n is large.

Let us now consider a prediction problem, where we want to predict the value of a new Y_0 with sample space $\mathcal{Y}$, whose distribution also depends on $\boldsymbol{\theta}$. We can always write $Y_0 = \varphi(\boldsymbol{\theta}, U)$, where U is a pivotal random variable with known distribution and sample space $\mathcal{U}$, and φ is a mapping from $\Theta \times \mathcal{U}$ to $\mathcal{Y}$ [12]. After observing the data $\boldsymbol{y}$, our knowledge about $\boldsymbol{\theta}$ is represented by the possibility distribution $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$. By Zadeh's extension principle [15], our knowledge of Y_0 conditionally on $U = u$ is, thus, represented by the possibility distribution $\pi_{Y_0|\boldsymbol{y},u} = \varphi(\pi_{\boldsymbol{\theta}|\boldsymbol{y}}, u)$ defined as

$$\pi_{Y_0|\boldsymbol{y},u}(y) = \sup_{\{\theta \in \Theta : \varphi(\theta,u)=y\}} \pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\theta) \tag{3}$$

for all $y \in \mathcal{Y}$. The mapping $\widetilde{Y} : [0,1] \to [0,1]^{\mathcal{Y}}$ such that $u \mapsto \pi_{Y_0|\boldsymbol{y},u}$ is, then, a RFS representing statistical evidence about Y_0.

3 Application to Regression Neural networks

We consider a neural network for regression with weight vector $\boldsymbol{w} \in \mathbb{R}^N$. The output for input x is denoted by $f(x; \boldsymbol{w})$. We assume that the response variable for an input vector x can be written as $Y = f(x; \boldsymbol{w}) + \sigma U$, where σ is the error standard deviation and $U \sim N(0,1)$ is a random variable with standard normal distribution. Given iid data $\{(x_i, y_i)\}_{i=1}^n$, the network is trained by maximizing the penalized log-likelihood

$$\ell_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = -n \log \sigma - \frac{1}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f(x_i; \boldsymbol{w}))^2 - \sum_{j=1}^{N} \lambda_j w_j^2, \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{w}^T, \sigma)^T$ is the vector of all parameters in the model and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$ is a vector of N regularization coefficients. The general form of the regularizer in (4) allows us to specify a distinct regularization coefficient for each weight; typically λ_j is set to 0 if w_j is a bias term. Our approach is based on a second-order approximation of the penalized log-likelihood (4) and a linear approximation of the map $\boldsymbol{w} \mapsto f(x; \boldsymbol{w})$ for a given input x. These two approximations are detailed below.

Possibility distribution of $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{w}}^T, \widehat{\sigma})^T$ be a global maximizer of $\ell_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$. We define the joint possibility distribution of $\boldsymbol{\theta}$ as $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta}) = \exp\left[\ell_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) - \ell_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})\right]$. We note that $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ is proportional to the product of the relative likelihood (1) and a GFV $\pi_0 \sim \mathrm{GFV}(\boldsymbol{0}, 2\,\mathrm{diag}(\boldsymbol{\lambda}))$, which can be seen as encoding prior information. Using the normal approximation (2), $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ can be approximated by a GFV with mode $\widehat{\boldsymbol{\theta}}$ and precision matrix $\mathcal{I}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}) = -\frac{\partial^2 \ell_{\boldsymbol{\lambda}}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$. Using simple calculations, it can be shown that

$$\mathcal{I}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} \boldsymbol{H} & \boldsymbol{v} \\ \boldsymbol{v}^T & a \end{pmatrix} \tag{5}$$

with

$$\boldsymbol{H} = -\left.\frac{\partial^2 \ell_{\boldsymbol{\lambda}}}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}, \quad \boldsymbol{v} = -\left.\frac{\partial^2 \ell_{\boldsymbol{\lambda}}}{\partial \boldsymbol{w} \partial \sigma}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = (4/\widehat{\sigma})\boldsymbol{\lambda} \odot \widehat{\boldsymbol{w}},$$

and $a = -\left.\frac{\partial^2 \ell_{\boldsymbol{\lambda}}}{\partial \sigma^2}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = 2n/\widehat{\sigma}^2$, where $\odot$ denotes pointwise multiplication.

Prediction. We now wish to predict a new outcome Y_0 of the response for $x = x_0$; it can be written as $Y_0 = f(x_0, \boldsymbol{w}) + \sigma U = \varphi(x_0, \boldsymbol{\theta}, U)$, with $U \sim N(0,1)$. Given $U = u$, the uncertainty Y_0 is constrained by the possibility distribution $\pi_{Y_0|\boldsymbol{y},u} = \varphi(x_0, \pi_{\boldsymbol{\theta}|\boldsymbol{y}}, u)$. This possibility distribution can be approximated by linearizing $f(x_0; \boldsymbol{w})$ around $\widehat{\boldsymbol{w}}$, which gives

$$f(x_0; \boldsymbol{w}) \approx f(x_0; \widehat{\boldsymbol{w}}) + \boldsymbol{g}(x_0)^T(\boldsymbol{w} - \widehat{\boldsymbol{w}}),$$

with $\boldsymbol{g}(x_0) = \left.\frac{\partial f(x_0;\boldsymbol{w})}{\partial \boldsymbol{w}}\right|_{\boldsymbol{w}=\widehat{\boldsymbol{w}}}$. With this approximation, we have

$$Y_0 \approx (\boldsymbol{g}(x_0)^T, U)\boldsymbol{\theta} + f(x_0, \widehat{\boldsymbol{w}}) - \boldsymbol{g}(x_0)^T\widehat{\boldsymbol{w}}.$$

From Proposition 1, assuming matrix $\mathcal{I}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})$ to be positive definite, possibility distribution $\pi_{Y_0|\boldsymbol{y},u}$ can then be approximated by a GFN with mode $f(x_0, \widehat{\boldsymbol{w}}) + \widehat{\sigma}u$ and precision

$$h(x_0, u) = \left[\begin{pmatrix} \boldsymbol{g}(x_0)^T & u \end{pmatrix} \mathcal{I}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})^{-1} \begin{pmatrix} \boldsymbol{g}(x_0) \\ u \end{pmatrix} \right]^{-1}. \tag{6}$$

The inverse of the precision matrix (5) can be written as

$$\mathcal{I}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})^{-1} = \frac{1}{c} \begin{pmatrix} c\boldsymbol{C}^{-1} & -\boldsymbol{H}^{-1}\boldsymbol{v} \\ -\boldsymbol{v}^T\boldsymbol{H}^{-1} & 1 \end{pmatrix},$$

where $\boldsymbol{C} = \boldsymbol{H} - \boldsymbol{v}\boldsymbol{v}^T/a = \boldsymbol{H} - 8(\boldsymbol{\lambda}\boldsymbol{\lambda}^T) \odot (\widehat{\boldsymbol{w}}\widehat{\boldsymbol{w}}^T)/n$, and

$$c = a - \boldsymbol{v}^T\boldsymbol{H}^{-1}\boldsymbol{v} = \frac{2n}{\widehat{\sigma}^2}\left(1 - \frac{8}{n}(\boldsymbol{\lambda}\boldsymbol{\lambda}^T) \odot (\widehat{\boldsymbol{w}}^T\boldsymbol{H}^{-1}\widehat{\boldsymbol{w}})\right).$$

Hence, (6) can be written as

$$\begin{aligned} h(x_0, u) &= c\left[\left\{c\boldsymbol{g}(x_0)^T\boldsymbol{C}^{-1} - \boldsymbol{v}^T\boldsymbol{H}^{-1}u\right\}\boldsymbol{g}(x_0) - \left[\boldsymbol{g}(x_0)^T\boldsymbol{H}^{-1}\boldsymbol{v} + u\right]u\right]^{-1} \\ &= \frac{1}{\alpha + \gamma u + u^2/c} \end{aligned} \tag{7}$$

with $\alpha = \boldsymbol{g}(x_0)^T\boldsymbol{C}^{-1}\boldsymbol{g}(x_0)$ and $\gamma = -2c^{-1}\boldsymbol{g}(x_0)^T\boldsymbol{H}^{-1}\boldsymbol{v}$. The predictive RFS is, thus, $\widetilde{Y}(x_0) : U \mapsto \mathrm{GFN}(f(x_0; \widehat{\boldsymbol{w}}) + U\widehat{\sigma}, h(x_0, U))$. We can observe that both the mode and the precision of $\widetilde{Y}(x_0)(U)$ depend on U: $\widetilde{Y}(x_0)$ is, thus, not a Gaussian random fuzzy number (GRFN) as defined in [7]. The degrees of belief and plausibility for any real interval can easily be computed by Monte Carlo simulation. Alternatively, we can observe that, for large n, the terms γu and $c^{-1}u^2$

become negligible compared to α in the denominator on the right-hand side of (7); replacing u and u^2 by their expectations, $h(x_0, u)$ can be approximated by $1/(\alpha + 1/c)$. RFS $\widetilde{Y}(x_0)$ is then, approximately, a GRFN with mean $f(x_0; \widehat{\boldsymbol{w}})$, variance $\widehat{\sigma}^2$ and precision $h(x_0) = 1/(\alpha + 1/c)$.

Remark 1 *In the above derivations, we have assumed that (i)* $\widehat{\boldsymbol{\theta}}$ *is a global maximizer of* $\ell_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$*, and (ii) precision matrix* (5) *is positive definite. The first assumption is necessary to ensure that the possibility distribution* $\pi_{\boldsymbol{\theta}|\boldsymbol{y}}$ *does not take values greater than one. It is very difficult, if not impossible, to guarantee that this assumption is verified, but we can ensure that we have reached a high enough maximum by running the optimization algorithm a large number of times. Assumption (ii) ensures that the inverse of the precision matrix exists and the precisions* (6) *are positive. As we will see in Section 4, this assumption is usually not verified exactly as matrix* $\mathcal{I}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})$ *typically has a small number of negative eigenvalues. If necessary, we may add a small quantity to the diagonal elements of* $\mathcal{I}_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})$ *to make it nonsingular and well conditioned.*

4 Simulation results

To evaluate the quality of the approximations performed in Section 3 and study some properties of the corresponding predictive belief functions, we considered the `Boston` dataset included in the R package `MASS`. We considered only three of the most informative predictors: `crim`, `zn` and `lstat`, which were normalized with zero mean and unit standard deviation. The data were split into a training set of size 300 and a test set of size 206. A network with one layer of 50 hidden units with Exponential Linear Unit (ELU) activation functions was fit to the data. The regularization coefficients had the same value $\lambda_j = \lambda$ for non-bias weights and $\lambda_j = 0$ for bias weights. Coefficient λ was determined by five-fold cross-validation, yielding $\lambda = 0.1$.

Figure 1 shows three examples of exact and approximated possibility distributions $\widetilde{Y}(x)(u)$ for different test input vectors x and random numbers u. The "exact" possibilities $\pi_{Y_0|\boldsymbol{y},u}(y)$ were computed by maximizing the log-likelihood $\ell_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ subject to the constraint $f(x, \boldsymbol{w}) = y$. As we can see, the exact possibility distributions are almost undistinguishable from their Gaussian approximations, which are themselves very well approximated by GRFNs.

The lower and upper predictive cdfs computed using the Gaussian approximation are shown in Figure 2, together with the GRFN approximation with fixed precision. Again, we can see that this latter approximation is excellent: the predictive RFSs are very well approximated by GRFNs. Figure 2 also displays the upper and lower predictive cdfs obtained by the ENNreg model [6].

Figure 3 shows calibration curves for the likelihood-belief functions introduced in this paper and for those computed by ENNreg. In [6], we defined calibration curves as plots of coverage probabilities of intervals centered on $\widehat{\mu}(x_0)$, with degree of belief α, for different values of $\alpha \in [0, 1]$; the predictive belief functions are calibrated if the curve is above the diagonal. In Figure 3, we display

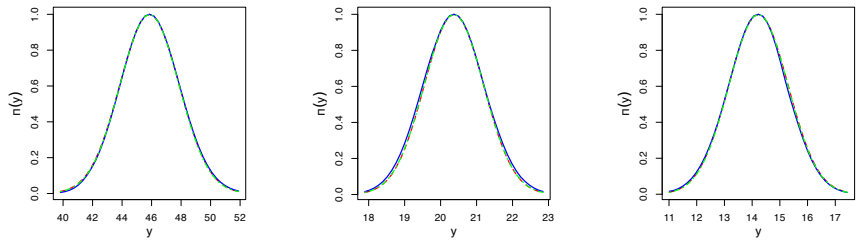


Fig. 1: Exact possibility distributions $\widetilde{Y}(x)(u)$ (solid blue lines), Gaussian approximations (red dashed lines) and Gaussian approximations with fixed precision (green dash-dotted lines) for three values of x and u.

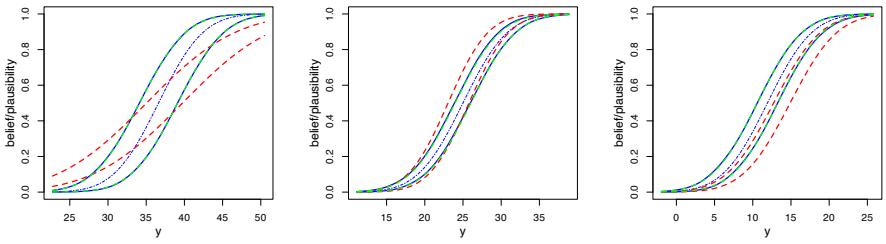
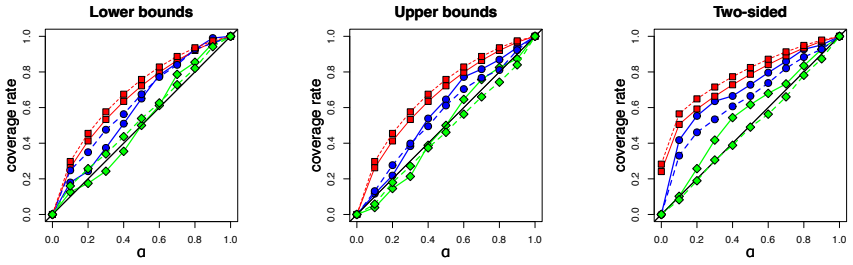


Fig. 2: Lower and upper cdfs of predictive RFSs $\widetilde{Y}(x)$ for the same three input vectors as those of Figure 1 (solid blue lines), Gaussian approximations with fixed precision (red dashed lines), and cdf of probabilistic prediction (blue dotted line). The predictive cdfs obtained by ENNreg are shown as cyan dash-dotted lines.

more detailed information in that we consider not only two-sided belief intervals centered at $\widehat{\mu}(x_0)$ (Figure 3c), but also one-sided intervals defined by a lower bound (Figure 3a) or an upper bound (Figure 3b). Furthermore, we display not only the coverage probabilities of different belief intervals, but also their average plausibilities.

We can see that, for both methods, the belief intervals are conservative (i.e., their coverage rates are greater than their belief degrees), and the coverage rates are bounded above by their average plausibilities, which corresponds to a stronger notion of calibration than that introduced in [6]. For this dataset, there appears to be little difference between the calibration graphs of the predictions obtained by two methods. As noted in [6], predictions can be adjusted using a validation sets to be as precise as possible, while remaining calibrated. A more extensive comparison between the two methods remains to be done.



(a) (b) (c)

Fig. 3: Calibration curves for belief lower bounds (a), belief upper bounds (b) and two-sided belief intervals (c) for likelihood-based belief functions (solid lines) and ENNreg (dashed lines). The lower green curves, middle blue curve and upper red curves correspond, respectively, to the coverage rates of probabilistic predictive intervals, coverage rates of belief intervals, and average plausibilities of belief intervals.

5 Conclusions

We have shown how to apply likelihood-based inference to regression neural network. The method is based on two approximations: the Gaussian approximation of the likelihood function, and the linearization of the network output with respect to the weights. These approximation make it possible to quantify prediction uncertainty by a RFS, which can itself be approximated by a Gaussian random fuzzy number as introduced in [7]. Experimental results with a real dataset suggest that these approximations are very accurate and that they allow us to compute calibrated predictive belief functions with low complexity (the most computationally expensive step being the calculation and inversion of the Hessian matrix, which need to be done only once). These preliminary results will need to be confirmed by much more extensive experiments. Also, in future work, our approach will be applied to a more realistic heteroscedastic model in which the conditional variance is a function of the inputs.

References

1. M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.
2. I. Couso and L. Sánchez. Upper and lower probabilities induced by a fuzzy random variable. *Fuzzy Sets and Systems*, 165(1):1–23, 2011.
3. T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

4. T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.
5. T. Denœux. An evidential neural network model for regression based on random fuzzy numbers. In S. Le Hégarat-Mascle, I. Bloch, and E. Aldea, editors, *Belief Functions: Theory and Applications*, pages 57–66, Cham, 2022. Springer International Publishing.
6. T. Denœux. Quantifying prediction uncertainty in regression using random fuzzy sets: the ENNreg model. *IEEE Transactions on Fuzzy Systems*, 31:3690–3699, 2023.
7. T. Denœux. Reasoning with fuzzy and uncertain evidence using epistemic random fuzzy sets: General framework and practical models. *Fuzzy Sets and Systems*, 453:1–36, 2023.
8. T. Denœux. Uncertainty quantification in logistic regression using random fuzzy sets and belief functions. *International Journal of Approximate Reasoning*, 168:109159, 2024.
9. T. Denœux, O. Kanjanatarakul, and S. Sriboonchitta. A new evidential k-nearest neighbor rule based on contextual discounting with partially supervised learning. *International Journal of Approximate Reasoning*, 113:287–302, 2019.
10. E. Hüllermeier and W. Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.
11. O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Forecasting using belief functions: an application to marketing econometrics. *International Journal of Approximate Reasoning*, 55(5):1113–1128, 2014.
12. O. Kanjanatarakul, S. Sriboonchitta, and T. Denœux. Prediction of future observations using belief functions: A likelihood-based approach. *International Journal of Approximate Reasoning*, 72:71–94, 2016.
13. H. T. Nguyen. On random sets and belief functions. *Journal of Mathematical Analysis and Applications*, 65:531–542, 1978.
14. D. A. Sprott. *Statistical Inference in Science*. Springer-Verlag, Berlin, 2000.
15. L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning –I. *Information Sciences*, 8:199–249, 1975.
16. L. A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.