Evidential Reasoning and Learning: a Survey

Federico Cerutti[1,2] , **Lance M. Kaplan**[3] and **Murat Şensoy**[4*]
[1]University of Brescia, Italy
[2]Cardiff University, UK
[3]US DEVCOM ARL, USA
[4]Amazon, Alexa AI, London, UK
federico.cerutti@unibs.it, lance.m.kaplan.civ@army.mil, msensoy@amazon.co.uk

Abstract

When collaborating with an artificial intelligence (AI) system, we need to assess when to trust its recommendations. Suppose we mistakenly trust it in regions where it is likely to err. In that case, catastrophic failures may occur, hence the need for Bayesian approaches for reasoning and learning to determine the confidence (or epistemic uncertainty) in the probabilities of the queried outcome. Pure Bayesian methods, however, suffer from high computational costs. To overcome them, we revert to efficient and effective approximations. In this paper, we focus on techniques that take the name of evidential reasoning and learning from the process of Bayesian update of given hypotheses based on additional evidence. This paper provides the reader with a gentle introduction to the area of investigation, the up-to-date research outcomes, and the open questions still left unanswered.

1 Introduction

Even in simple collaboration scenarios—like those in which an artificial intelligence (AI) system assists a human operator with predictions—the human has developed insights (i.e., a mental model) of when to trust the AI system with its recommendations [Bansal *et al.*, 2019b]. If the human mistakenly trusts the AI system in regions where it is likely to err, catastrophic failures may occur. This is a strong argument favouring Bayesian approaches to probabilistic reasoning: research in the intersection of AI and human-computer interaction (HCI) has found that interaction improves when setting expectations right about what the system can do and how well it performs [Kocielnik *et al.*, 2019; Bansal *et al.*, 2019a]. Guidelines have been produced [Amershi *et al.*, 2019], and they recommend to *Make clear what the system can do*, and *Make clear how well the system can do what it can do*.

To identify such regions where the AI system is likely to err, we need to distinguish between (at least) two different sources of uncertainty: *aleatory* (or *aleatoric*) and *epistemic* uncertainty [Hora, 1996; Hüllermeier and Waegeman, 2021]. Aleatory uncertainty refers to the variability in the outcome of an experiment due to inherently random effects (e.g. flipping a fair coin): no additional source of information but Laplace's daemon[1] can reduce such variability. Epistemic uncertainty refers to the epistemic state of the agent using the model, hence its lack of knowledge that—in principle—can be reduced on the basis of additional data samples.

This paper dwells on the research at the intersection of quantifying aleatory and epistemic uncertainty in reasoning and learning, while using very efficient approximations based upon the idea of updating the Bayesian posterior in light of further evidence collected in favour (or against) a hypothesis. We primarily focus on the case of uncertain probabilities represented as beta or Dirichlet distributions following the Bayesian statistics paradigm (Section 2).

Unlike existing surveys on approaches for quantifying epistemic uncertainty in (deep) learning, e.g., [Hüllermeier and Waegeman, 2021; Abdar *et al.*, 2021], in this paper, we aim at giving a overview of the challenges associated with the reasoning in the presence of epistemic uncertainty and with learning both with full and partial data. Logical reasoning in the presence of aleatory and epistemic uncertainty (Section 3) brings entirely novel problems that need to be addressed when wishing to limit the need for computational resources. Evidential reasoning thus introduces the idea of choosing either beta or Dirichlet distributions to represent uncertain probabilities and then using efficient methods—such as the the moment matching—for manipulating them. We illustrate this idea using Cerutti *et al.*'s proposal [2022] as it builds on the notion of probabilistic circuits [Choi *et al.*, 2020], which can encompass a large set of reasoning problems. It can be used as a common representation framework for various other approaches, from Bayesian networks [Darwiche, 2009]—an interested reader here is also referred to Rohmer's survey [2020] on uncertainties in Bayesian networks—to probabilistic logic programming [Fierens *et al.*, 2015].

We further discuss the challenges of ascertaining epistemic and aleatory uncertainty of probabilistic circuits parameters, particularly with partial observability of the training data (Section 4). This is the most challenging and the least devel-

[*] The work was done prior to joining Amazon.

[1] "An intelligence that, at a given instant, could comprehend all the forces by which nature is animated and the respective situation of the beings that make it up" [Laplace, 1825, p.2].

oped task, and we illustrate the preliminary results achieved so far. We also add a few remarks on the importance of epistemic uncertainty in determining the logical structures underpinning our reasoning.

We finally discuss how to ascertain uncertain probabilities from the real world (Section 5). Unsurprisingly, they are either provided by an oracle (e.g., an intelligence analyst) or learnt from raw data. We illustrate approaches to evidential learning, mainly focusing on Şensoy *et al.*'s proposal [2018] in the light of its simplicity. Instead of assigning probabilities to the classes, Şensoy *et al. count the pieces of evidence* in favour of each of the possible classes. Then, some of them also add a regularisation term to the loss function to ensure that the total number of pieces of evidence should shrink towards zero—equivalent to say: *I do not know*—when a data sample cannot be correctly classified.

We conclude the paper (Section 6) with a discussion on open questions still left unanswered.

2 A Primer in Bayesian Statistics

Given a learning model—e.g. a neural network—whose parameters $\boldsymbol{w}$ we want to learn from a dataset $\mathcal{D}$, the *frequentist paradigm* to statistics considers $\boldsymbol{w}$ a fixed parameter and searches for an estimation. A widely used estimator is the *maximum likelihood* in which $\boldsymbol{w}$ is set to the value that maximises $p(\mathcal{D} \mid \boldsymbol{w})$. The negative log of the likelihood function is often chosen as the loss function.

The *Bayesian paradigm*, instead, considers that the observed data set $\mathcal{D}$ tightens probabilistic knowledge about the value of $\boldsymbol{w}$. Bayes theorem is used to convert a prior probability into a posterior probability by incorporating the evidence provided by the observed data. Given the parameters of our model $\boldsymbol{w}$, we can capture our assumptions about $\boldsymbol{w}$, before observing the data, in the form of a prior probability distribution $p(\boldsymbol{w})$. The effect of the observed data $\mathcal{D}$ is expressed through the conditional $p(\mathcal{D} \mid \boldsymbol{w})$, hence Bayes theorem takes the form:

$$p(\boldsymbol{w} \mid \mathcal{D}) = \frac{\overbrace{p(\mathcal{D} \mid \boldsymbol{w})}^{\text{likelihood}}\overbrace{p(\boldsymbol{w})}^{\text{prior}}}{p(\mathcal{D})} \tag{1}$$

The denominator in (1) is the normalisation constant, which ensures that the posterior distribution on the left-hand side is a valid probability density and integrates to one. If the posterior distribution is in the same probability distribution family as the prior probability distribution, the prior is called a *conjugate prior* for the likelihood. Choosing a conjugate prior often leads to a closed-form expression of the posterior, thus avoiding the need for a numerical integration for computing the denominator of (1).

2.1 Binary Classification and Uncertain Probabilities

When facing a binary classification a *complete dataset* $\mathcal{D}$ is then a sequence (allowing for repetitions) of examples, each of those is a vector of instantiations of independent Bernoulli distributions with true but unknown parameter π.

From this, the *likelihood* is thus: $p(\mathcal{D} \mid \pi) = \prod_{n=1}^{|\mathcal{D}|} p(x_n \mid \pi) = \prod_{n=1}^{N} \pi^{x_n}(1-\pi)^{1-x_n}$ where x_i represents the i-th example in the dataset $\mathcal{D}$, that is assumed to hold either the value 1 or 0.

To develop a Bayesian analysis of the phenomenon, we can choose as prior the beta distribution, with parameters $\boldsymbol{\alpha} = \langle \alpha_x, \alpha_{\overline{x}} \rangle$, $\alpha_x > 0$ and $\alpha_{\overline{x}} > 0$, that is conjugate to the Bernoulli:

$\mathrm{Beta}(\pi \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_x+\alpha_{\overline{x}})}{\Gamma(\alpha_x)\Gamma(\alpha_{\overline{x}})}\pi^{\alpha_x-1}(1-\pi)^{\alpha_{\overline{x}}-1}$ where $\Gamma(t) \equiv \int_0^\infty u^{t-1}e^{-u}\mathrm{d}u$ is the *gamma* function.

Considering a beta distributed prior $\mathrm{Beta}(\pi \mid \widehat{\boldsymbol{\alpha}})$ and the Bernoulli likelihood function, and given $|\mathcal{D}|$ observations $\boldsymbol{m} = \langle m_x, m_{\overline{x}} \rangle$ of x, viz., m_x observations of $x = 1$, $m_{\overline{x}}$ observations of $x = 0$, and $m_x + m_{\overline{x}} = |\mathcal{D}|$:

$$p(\pi \mid \mathcal{D}, \widehat{\boldsymbol{\alpha}}) = \mathrm{Beta}(\pi \mid \widehat{\boldsymbol{\alpha}} + \boldsymbol{m}) \tag{2}$$

Thus, the parameters of a beta distribution can be considered pseudocounts [Murphy, 2012] of *pieces of evidence* for the two outcomes of a phenomenon, and the beta distribution itself can be seen as a representation of the uncertain probability associated with the phenomenon. Among the various priors, using $\widehat{\boldsymbol{\alpha}} = \mathbf{1} = \langle 1, 1 \rangle$ is equivalent to using the uniform distribution, which represents a non-informative prior that maximises entropy.

2.2 Multi-class Classification

The *Dirichlet* distribution generalises the beta distribution to K dimensions: indeed, the marginals of a Dirichlet distribution are beta distributions. $\mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}$ such that $\sum_k \pi_k = 1$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, $\alpha_k > 0$ and $\alpha_0 = \sum_{k=1}^{K}\alpha_k$.

Considering a Dirichlet distribution prior and the categorical likelihood function—which is the generalisation of the Bernoulli to K dimensions—and considering $|\mathcal{D}|$ observations $\boldsymbol{m}$, the posterior when choosing as prior $\mathrm{Dir}(\boldsymbol{\pi} \mid \widehat{\boldsymbol{\alpha}})$ is then:

$$p(\boldsymbol{\pi} \mid \mathcal{D}, \widehat{\boldsymbol{\alpha}}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \widehat{\boldsymbol{\alpha}} + \boldsymbol{m}) \tag{3}$$

The uniform prior is given by $\mathrm{Dir}(\boldsymbol{\pi} \mid \mathbf{1})$.

3 Evidential Reasoning

Following Kimmig *et al.*'s approach [2017], let us consider a propositional logic theory $\mathcal{T}$ over a set of variables $\mathcal{V}$. An interpretation of $\mathcal{V}$ assigns a truth value from the set $\{\top, \bot\}$ to every variable in $\mathcal{V}$. The set $\mathcal{M}(\mathcal{T})$ of models of theory $\mathcal{T}$ contains exactly those interpretations of $\mathcal{V}$ for which $\mathcal{T}$ evaluates to true. Given the set of literal $\mathcal{L}$ for the variables in $\mathcal{V}$, let $p : \mathcal{L} \to [0, 1]$ and, for $l \in \mathcal{L}$, $p(l) + p(\neg l) = 1$.

Given a query $q \subseteq \mathcal{L}$ and $\mathcal{I}(q) = \{I \mid I \in \mathcal{M}(\mathcal{T}) \wedge q \subseteq I\}$ the set of interpretations where the query is true, then the probabilistic inference task is:

$$\textbf{PROB}(q) = \sum_{I \in \mathcal{I}(q)} \prod_{l \in I} p(l). \tag{4}$$

The complexity of probabilistic reasoning in (4) is hidden in the computation of $\mathcal{I}(q)$ which is $\#P$-complete [Valiant,

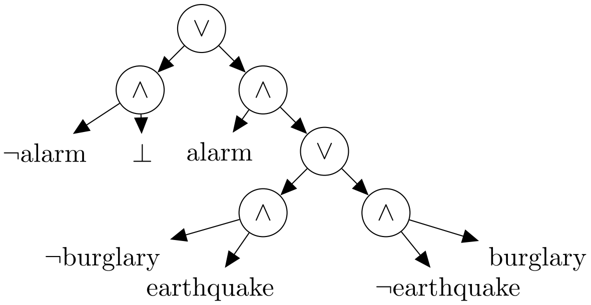

Figure 1: A compiled circuit of (5) allowing for polytime calculation of its set of models.

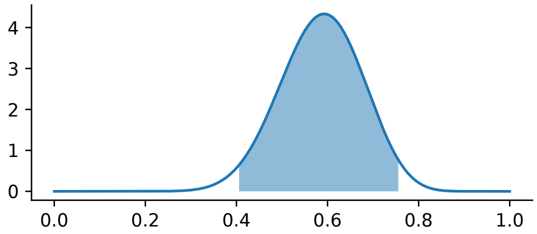

Figure 2: $\mathrm{Beta}(17, 12)$: in shaded blue the 95% confidence interval.

1979a; Valiant, 1979b]. However, with knowledge compilation techniques [Darwiche and Marquis, 2002], a propositional theory can be compiled off-line into a rich, nested language based on representing propositional sentences using directed acyclic graphs or *circuit*—where each leaf is labelled with elements of $\mathcal{L}$ or $\{\top, \bot\}$, and each internal node is labelled either $\wedge$ or $\vee$—which is then used on-line to answer a large number of queries in polynomial time.

Let us revisit a simplified version of the alarm-burglary-earthquake notorious example [Kim and Pearl, 1983]: an alarm goes off and that can be triggered only by either a burglary or an earthquake. In propositional logic:

$$\text{alarm} \wedge (\text{alarm} \iff (\text{burglary} \vee \text{earthquake})\ ). \qquad (5)$$

Figure 1 depicts a compiled circuit of (5) allowing for polynomial-time calculation of its set of models and the interpretation of a query.

Figure 1 can be transformed into an equivalent probabilistic circuit by: changing each $\wedge$-labelled node into a multiplication; each $\vee$-labelled node into an addition; and considering the probabilistic labels associated with each literal at the leaves. In general, probabilistic circuits (PCs) [Choi *et al.*, 2020] refer to a family of tractable probabilistic models—which include, among others, arithmetic circuits [Darwiche, 2003], probabilistic sentential decision diagrams [Kisa *et al.*, 2014], and sum-product networks [Poon and Domingos, 2011]—that are known to be able to closely capture the probability space in density estimation tasks [Dang *et al.*, 2022; Peharz *et al.*, 2020], usually under the independence assumption, while allowing tractable inference of many useful queries.

From Section 2.1, we see that uncertain probabilities can be expressed in the form of beta distributions. Figure 2 depicts the PDF of $\mathrm{Beta}(17, 12)$, whose expected value is 0.59 and variance is $8.06 \cdot 10^{-3}$ thus representing an imprecise probability around 0.6. The research question then becomes: *how to reason efficiently when dealing with imprecise probabilities in a probabilistic circuit?*.

Cerutti *et al.* [2019] label leaves of probabilistic circuits—built from aProbLog programs [Kimmig *et al.*, 2011]—with beta distributions and propose *multiplication* and *addition* operators that receive as input two beta distributions, and return a beta distribution that matches the first two moments of the distribution resulting from the multiplication (resp. addition) of the two input distributions. Beta distributions have two parameters, hence having the first two moments—the expected value and the variance—suffices for determining the two parameters of a beta distribution with the same expected value and the same variance—modulo certain values that are impossible for a beta distribution.

Cerutti *et al.* [2022] then generalise the idea by proposing an algorithm for computing the probabilistic (conditional) inferences with imprecise probabilities that accepts as input the covariance between the various distributions associated with the leaves of the circuit. It also no longer requires circuits built from aProbLog programs.

Probabilistic circuits are known to be able to solve also the problem of inference in Bayesian networks [Bacchus *et al.*, 2009]. That might induce the computational overhead of deriving a probabilistic circuit from a Bayesian network, while more efficient algorithms exist. Kaplan and Ivanovska [2018] use the moment-matching approach to deal with inferences in singly-connected Bayesian networks using a modified version of the message-passing algorithm for belief propagation.

Subjective logic [Jøsang, 2016] provides (1) an alternative, intuitive way of representing the parameters of beta-distributed random variables, and (2) a set of operators for manipulating them. Unlike previously discussed approaches [Kaplan and Ivanovska, 2018; Cerutti *et al.*, 2019; Cerutti *et al.*, 2022], subjective logic approximates Bayesian reasoning via a *least commitment principle*, i.e., matching the expected values but then maximising the variance.

Subjective logic also provides a mapping with Dempster–Shafer theory [Dempster, 1967; Shafer, 1976], which abandons the additivity principle of probability theory, viz. that the sum of probabilities on all pairwise exclusive possibilities must add up to one. In this way, the lack of evidence to support any specific probability can be explicitly expressed by assigning belief mass to the whole frame of discernment, which comprises the set of exclusives possible states. Smets introduced [1993] a computationally efficient method to manipulate Dempster-Shafer belief assignments.

Figure 3 compares the proposals discussed above—and a simple Monte Carlo approach over 100 samples—when facing the problem of probabilistic inference over a singly-connected Bayesian network [Cerutti *et al.*, 2022, Fig. 12a]. We judge the quality of the results on how well their expressions of uncertainty capture the spread between its projected probability and the actual ground truth probability [Kaplan and Ivanovska, 2018]. By knowing the ground truth, confidence bounds can be formed around the projected probabilities at a significance level and determine the fraction of cases

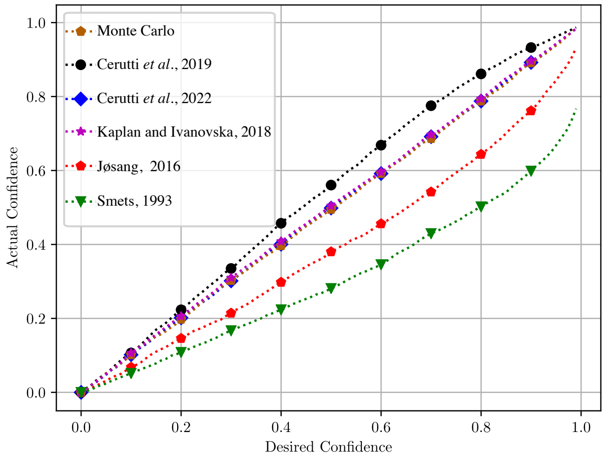

Figure 3: Actual versus desired significance of bounds derived from the uncertainty when dealing with probabilistic inference over the single-connected Bayesian network [Cerutti *et al.*, 2022, Fig. 12a]. Monte Carlo has been run over 100 samples. Best closest to the diagonal: below the diagonal, the desired confidence is greater than the actual confidence and thus such an approach is generally overconfident in its assessment of epistemic uncertainty.

when the ground truth falls within the bounds. If the uncertainty is well determined, this fraction should correspond to the strength of the confidence interval [Kaplan and Ivanovska, 2018, Appendix C]; thus best results are closest to the diagonal. From Figure 3, it is immediate to see that: (1) the Monte Carlo approach, [Cerutti *et al.*, 2022], and [Kaplan and Ivanovska, 2018] comfortably sit on the diagonal; (2) [Cerutti *et al.*, 2019] is underconfident in its evaluation, which is due to the independence assumption; and (3) both [Jøsang, 2016] and [Smets, 1993] are overconfident.

4 Evidential Learning of Parameters (and Structures) of Probabilistic Circuits

For each of the leaves of probabilistic circuits—such as the one that can be derived from Figure 1—that is labelled with a literal, there is an associated distribution. When a set of complete observations is given, i.e., a set of interpretations where each interpretation assigns a truth value to each literal, learning the associated distributions $\boldsymbol{\theta}$ is a relatively simple task resulting in counting the pieces of evidence for the various outcomes considered by the distributions and use them to update a chosen prior (see Sections 2.1 and 2.2).

Instead, when learning with incomplete observations, a.k.a. the incomplete-features problem, traditional approaches include skipping missing values or performing data reconstruction (data completion) before using the data in a model. This, however, might fail to capture the joint distributions of the variables faithfully. The expectation-maximization (EM) framework has classically been used, for instance, to learn conditional probabilities for Bayesian networks with incomplete training data, e.g., [Lauritzen, 1995].

There is, however, limited work aimed at identifying the posterior distributions learned from incomplete training data, particularly when it comes to determining their covariances.

One of the most prominent approaches is the Online Bayesian Moment Matching (BMM) [Rashwan *et al.*, 2016], which approximates the posterior distribution as a product of Dirichlet random variables, presuming that the group of conditional probabilities are statistically independent. It is known that the ground-truth, i.e., latent distribution, of $\boldsymbol{\theta}$ is Dirichlet, so we begin with a prior that is a product of Dirichlets with respect to the weights of each sum node in a probabilistic circuit.

The evaluation of a given circuit consists of alternating sums and products, which means that the posterior becomes a mixture of products of Dirichlet distributions. While the mixture of Dirichlet products admits a closed form expression for its posterior distribution, unfortunately it is also computationally intractable since the number of mixture components is exponential in the number of sum nodes in the circuit [Rashwan *et al.*, 2016]. BMM solves this problem by presuming the posterior is a product of Dirichlets which is fit via moment matching.

Alternatively, estimating the posterior distribution could be addressed in a two-step procedure: (1) first estimating the expected value via, for instance, EM; and (2) then estimating the covariance of parameters using the Fisher information matrix (FIM) [Ly *et al.*, 2017] as per the Bernstein-von Mises theorem [van der Vaart, 1998]. Suppose to have $\boldsymbol{T}$ uninstantiated observations, then the FIM is given by:

$$\mathbb{E}_{p(\boldsymbol{T}|\boldsymbol{\theta})}[\nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{T} \mid \boldsymbol{\theta})) \cdot (\nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{T} \mid \boldsymbol{\theta})))^{\mathrm{T}}]$$

where $\nabla_{\boldsymbol{\theta}} \log(p(\boldsymbol{T} \mid \boldsymbol{\theta}))$, the *score function*, is the gradient of log likelihood function, and FIM is the covariance of score function which provides an assessment of the certainty of the model. The Bernstein-von Mises theorem [van der Vaart, 1998] then informs us that, under certain conditions, posterior distributions converge to normal distributions centred at the maximum likelihood estimator with covariance matrix given by the inverse of the FIM.

For incomplete training data, Kaplan *et al.* [2020] provide a derivation for the FIM (EM-FIM), which, having also the interpretation of being the negative expected Hessian of score function [Martens, 2020], can be estimated using a Gaussian approximation of the parameters (EM-GA). Hougen *et al.* compared BMM, EM-FIM, and EM-GA [2021], and suggest that, when data not only is incomplete but is scarce too, best approximations are provided by EM-FIM and BMM, which seems also to be the best option when the datasets are becoming more complete but independent. When inducing stronger statistical dependencies between variables, unsurprisingly BMM's performance deteriorates while EM-FIM in particular still appears to provide better estimations. Unfortunately, EM-FIM appears also to be the slowest and the least scalable of the alternatives.

Finally, although so far the circuit structures has been assumed to be given, there are, however, several algorithms for learning them, e.g., [Benjumeda *et al.*, 2019b; Benjumeda *et al.*, 2019a; Dang *et al.*, 2022]. Most of them rely on some score function that is used to guide a search procedure in the space of alternative structures. To our knowledge, the only evidential score function so far employed has been illustrated

[Cunnington *et al.*, 2021] in the specific case of learning logic programs—from which it is possible to derive probabilistic circuits following knowledge compilation techniques similar to the one summarised in Section 3—from positive and negative examples with an associated epistemic uncertainty in the form of Dirichlet distributions which are learnt using the evidential deep learning (EDL) approach we discuss in the following section.

5 Ascertain Evidence from the the Real World

The scientific and the intelligence communities have long relied on scales for classifying pieces of information in terms of both likeability (e.g., *likely*, *very likely*, ...) and confidence (e.g., *low confidence*, *medium confidence*, ...) [Mastrandrea *et al.*, 2011]. Following, for instance, a mapping similar to the one proposed by Jøsang [2016, Section 3.7.2], each combination of such statements can be represented by a proposition, for instance, with an associated beta distribution with low expected value (*unlikely*) and with large variance (*low confidence*).

The simplest case for learning uncertain probabilities from raw data is to use classifiers that expand on discriminative classifiers (e.g. logistic regression, vanilla neural network classifiers) [Murphy, 2012], which are models of the form $p(y \mid \boldsymbol{x})$ where a distribution over the K possible categories is estimated directly. Using the EDL (Evidential Deep Learning) approach [Şensoy *et al.*, 2018], the learning problem shifts from learning a probability distribution between the labels, to count pieces of evidence for each class. While a typical discriminative classifier makes a point estimate of $\boldsymbol{\pi}$ directly, EDL estimates a Dirichlet distribution over $\boldsymbol{\pi}$.

To illustrate the benefits of Şensoy *et al.*'s approach [2018], let us use the idea of rotating an MNIST—a database of handwritten digits [Lecun *et al.*, 1998]—image for the digit 1 to analyse the behaviour in the light of unknown inputs [Gal and Ghahramani, 2016; Louizos and Welling, 2017]. A softmax approach will identify a class as very likely (see top part of Figure 4) independently of the rotation angle. EDL [Şensoy *et al.*, 2018], instead, will raise the epistemic uncertainty (middle part of Figure 4) while reverting each class probability to the prior uniform distribution when the rotated digit stops resembling anything the machine has previously seen.

Each of the loss functions introduced by Şensoy *et al.* [2018] has two components: one aims at minimising the prediction error, the other the number of pieces of evidence generated for each class, thus learning to say *I do not known* when facing ambiguous datapoints. Concerning the first component $\mathcal{L}_i(\theta)$, the simplest—yet very effective—version looks at the prediction error in terms of the sum of squares loss $||\mathbf{y}_i - \boldsymbol{\pi}_i||_2^2$ which aims to achieve the joint goal of minimising the prediction error and the variance of the Dirichlet experiment generated by the neural net, specifically for each sample in the training set.

Regarding the second component of the loss function, EDL [Şensoy *et al.*, 2018] enforces a total count of evidence equal to zero for a sample if it cannot be correctly classified. In this case, the posterior Dirichlet distribution is equivalent to the prior distribution: when choosing a uniform prior distribution, this correspond to total uncertainty (maximum entropy). This aim can be achieved by incorporating a Kullback-Leibler (KL) divergence term that penalises cases that do not contribute to the data fit. The loss with this regularising term, in the case of a multi-epoch training procedure, reads

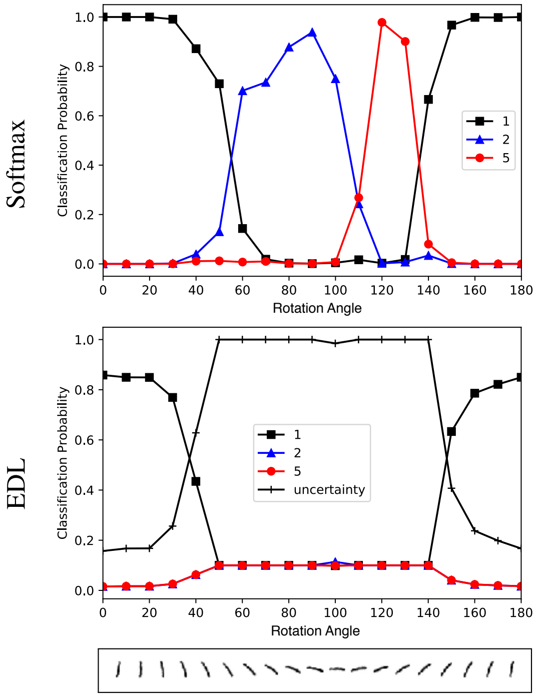

Figure 4: Classification of the rotated digit 1 (at bottom) at different angles between 0 and 180 degrees. **Top:** The classification probability is calculated using the *softmax* function considering three possible classes, digit 1, digit 2, and digit 5. **Middle:** The classification probability (aleatory uncertainty) considering the same three possible classes, and (epistemic) uncertainty are calculated using EDL [Şensoy *et al.*, 2018].

$$\begin{aligned}\mathcal{L}(\theta) = & \sum_{i=1}^{N} \mathcal{L}_i(\theta) + \\ & + \lambda_t \sum_{i=1}^{N} \mathrm{KL}[\mathrm{Dir}(\boldsymbol{\pi}_i \mid \tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}) || \mathrm{Dir}(\boldsymbol{\pi}_i \mid \mathbf{1})],\end{aligned} \tag{6}$$

where $\lambda_t = \min(1.0, t/10) \in [0, 1]$ is the annealing coefficient, t is the index of the current training epoch, $\mathrm{Dir}(\boldsymbol{\pi}_i \mid \mathbf{1})$ is the uniform Dirichlet distribution with $\mathbf{1} = [1, \ldots, 1]$, and lastly $\tilde{\boldsymbol{\alpha}}_i = \mathbf{y}_i + (1 - \mathbf{y}_i) \odot \boldsymbol{\alpha}_i$ is the Dirichlet parameters after removal of the non-misleading evidence from predicted parameters $\boldsymbol{\alpha}_i$ for sample i. The role of the annealing coefficient λ_t is paramount: by gradually increasing the effect of the KL divergence on the loss, the learning algorithm explores the parameter space and avoid premature convergence to the uniform distribution for the misclassified samples.

EDL [Şensoy *et al.*, 2018] is not the only proposal for ascertain epistemic uncertainty from raw data, nowadays a very florid research area as testified also by recent, detailed surveys [Hüllermeier and Waegeman, 2021; Abdar *et al.*, 2021]. Prior Networks [Malinin and Gales, 2018] also predicts Dirichlet distribution for classification. To avoid overconfident predictions, it uses an auxiliary data set as the out-of-distribution samples and explicitly trains the neural networks to give highly uncertain output for them. In their followup work, Malinin and Gales [2019] use loss functions similar to the ones introduced by Şensoy *et al.* [2018].

To overcome some of the limitations of Prior Networks, Charpentier *et al.* [2020] propose Posterior Networks which uses normalising flow [Rezende and Mohamed, 2015] for learning a latent representation of the input. They then learn a mapping to a Dirichlet distribution using a Bayesian loss function composed by (1) the Uncertain Cross Entropy loss introduced by Biloš *et al.* [2019], which increases confidence for observed data, and (2) a regulariser which favours smooth distributions. Kopetzki *et al.* analysed the relative performance of such proposed approaches [2021] focusing on the detection of adversarial attacks.

Also Gast and Roth [2018] suggest probabilistic output layers for classification (and regression) that require only minimal changes to existing networks. In particular, classification tasks can be approached with a Dirichlet layer that can be trained by conditional likelihood maximisation.

Haussmann *et al.* [2021] build upon the EDL proposal [Şensoy *et al.*, 2018] by converting it to a Bayesian neural network [Mackay, 1995]. To overcome the prohibitively large number of hyperparameters, Haussmann *et al.* derive a vacuous PAC [McAllester, 2003] bound that comprises the marginal likelihood of the predictor and a complexity penalty.

Building upon the notion of Knowledge as Justified Belief from the field of epistemology [Ichikawa and Steup, 2018], Virani *et al.* propose [2020] epistemic classifiers that use contextual information based on location of training data points in input and hidden layers to add reliability on individual predictions. Evidence to construct justification is gathered using various domain-agnostic neighbourhood operators. Bhushan *et al.* further developed the approach [2020] basing the classification upon a latent representation obtained using variational auto encoders [Kingma and Welling, 2014].

Variational auto encoders are also exploited by Şensoy *et al.* [2020], where the original EDL proposal [Şensoy *et al.*, 2018] has been extended so to harvest the same benefits there would be by using an auxiliary dataset of out-of-distribution samples, but without the costs of selection and creation, by using variational autoencoders and generative adversarial networks are incorporated to automatically generate out-of-distribution exemplars for training. Şensoy *et al.* [2020] demonstrate how that provides excellent estimates of uncertainty for in- and out-of-distribution samples, and adversarial examples on well-known data sets.

Finally, also approaches based upon the Dempster–Shafer theory [Dempster, 1967; Shafer, 1976] has been proposed. For instance, Denœux proposes [2019] to convert inputs (or higher-level features) into Dempster-Shafer mass functions and aggregating them by Dempster's rule of combination.

6 Conclusion

This paper dwells on the research at the intersection of quantifying aleatory and epistemic uncertainty in reasoning and learning, while using very efficient approximations based upon the idea of updating the Bayesian posterior in light of further evidence collected in favour (or against) a hypothesis. Evidential reasoning (Section 3) introduces the idea of choosing either beta or Dirichlet distributions to represent uncertain probabilities and then using efficient methods—such as the the moment matching—for manipulating them. Reasoning structures can be captured by probabilistic circuits: we discussed (Section 4) the challenges of ascertaining epistemic and aleatory uncertainty of their parameters. We finally discussed (Section 5) how to fathom uncertain probabilities from the real world, with large emphasis on the case where they are learnt from raw data. As a corollary here, it is worth mentioning that Amini *et al.* propose [2020] an evidential approach to regression. Also evidential learning approaches, in various forms, have been employed in a large variety of application domains, from chest radiography classification to cyber-threat classification.

Despite the large set of results, several research questions remain open. As discussed in Section 4, we are still far from having a coherent picture of the best algorithms for estimating aleatory and epistemic uncertainty of parameters in probabilistic circuits, leaving aside efficient—and user-friendly—implementations. Similarly, the role of epistemic uncertainty in parameter learning is far from being exhaustively eviscerated. In addition, Sections 3 and 4 focus on logical reasoning over propositional theories. Open research questions include how to effectively represent and reason about more complex cases, such as uncertain spatial and temporal relationships, which are topics of great practical importance too when considering, for instance, information retrieval or autonomous navigation.

Moreover, when dealing with real-world problems it is still unclear how to deal with an input which is classified with high epistemic uncertainty: does it identify a new class? For instance, Bao *et al.* use [2021] evidential learning to boost the performance of existing models to recognise actions in an open testing set. However, we believe further investigations are still needed, particularly when uncertainty might be linked to ambiguity. For instance, it might be advantageous for a classifier trained to distinguish between cats, dogs, and wolves, to be able to assign belief masses to ambiguous classes such as “this input is either a dog or a wolf, but not a cat.”

Finally, currently there are only very preliminary proposals [Cunnington *et al.*, 2021] trying to link evidential deep networks with evidential probabilistic circuits in a single neuro-symbolic architecture, leaving aside the possibility to use evidential reasoning in neuro-programming architectures. This is clearly an exciting research areas that, we are certain, will receive great attention in the near future.

References

[Abdar *et al.*, 2021] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. Rajendra Acharya, V. Makarenkov, and S. Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[Amershi *et al.*, 2019] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz. Guidelines for human-AI interaction. In *CHI*, pages 1–13, 2019.

[Amini *et al.*, 2020] A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep Evidential Regression. In *NeurIPS*, pages 14927–14937, 2020.

[Bacchus *et al.*, 2009] F. Bacchus, S. Dalmao, and T. Pitassi. Solving #SAT and Bayesian inference with backtracking search. *JAIR*, 34:391–442, 2009.

[Bansal *et al.*, 2019a] F. Bansal, B. Nushi, E. Kamar, W. Lasecki, D. Weld, and E. Horvitz. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *HCOMP*, pages 2–11, 2019.

[Bansal *et al.*, 2019b] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *AAAI*, pages 2429–2437, 2019.

[Bao *et al.*, 2021] Wentao Bao, Qi Yu, and Yu Kong. Evidential Deep Learning for Open Set Action Recognition. In *ICCV*, pages 13349–13358, 2021.

[Benjumeda *et al.*, 2019a] M. Benjumeda, C. Bielza, and P. Larrañaga. Learning tractable Bayesian networks in the space of elimination orders. *Artif. Intell.*, 274:66–90, 2019.

[Benjumeda *et al.*, 2019b] M. Benjumeda, S. Luengo-Sanchez, P. Larrañaga, and C. Bielza. Tractable learning of Bayesian networks from partially observed data. *Pattern Recogn.*, 91:190–199, 2019.

[Bhushan *et al.*, 2020] C. Bhushan, Z. Yang, N. Virani, and N. Iyer. Variational Encoder-Based Reliable Classification. In *ICIP*, pages 1941–1945, 2020.

[Biloš *et al.*, 2019] M. Biloš, B. Charpentier, and S. Günnemann. Uncertainty on Asynchronous Time Event Prediction. In *NeurIPS*, 2019.

[Cerutti *et al.*, 2019] F. Cerutti, L. M. Kaplan, A. Kimmig, and M. Şensoy. Probabilistic Logic Programming with Beta-Distributed Random Variables. In *AAAI*, pages 7769–7776, 2019.

[Cerutti *et al.*, 2022] F. Cerutti, L. M. Kaplan, A. Kimmig, and M. Şensoy. Handling Epistemic and Aleatory Uncertainties in Probabilistic Circuits. *Mach. Learn.*, 2022.

[Charpentier *et al.*, 2020] B. Charpentier, D. Zügner, and S. Günnemann. Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts. In *NeurIPS*, pages 1356–1367, 2020.

[Choi *et al.*, 2020] Y. Choi, A. Vergari, and G. Van den Broeck. Probabilistic circuits: A unifying framework for tractable probabilistic models. Technical report, 2020.

[Cunnington *et al.*, 2021] D. Cunnington, M. Law, A. Russo, J. Lobo, and L. M. Kaplan. Towards Neural-Symbolic Learning to support Human-Agent Operations. In *FUSION*, 2021.

[Dang *et al.*, 2022] M. Dang, A. Vergari, and G. Van den Broeck. Strudel: A fast and accurate learner of structured-decomposable probabilistic circuits. *IJAR*, 140:92–115, 2022.

[Darwiche and Marquis, 2002] Adnan Darwiche and Pierre Marquis. A Knowledge Compilation Map. *J. Artif. Int. Res.*, 17(1):229–264, September 2002.

[Darwiche, 2003] A. Darwiche. A differential approach to inference in Bayesian networks. *Journal of the ACM*, 50(3):280–305, 2003.

[Darwiche, 2009] A. Darwiche. *Modeling and Reasoning with Bayesian Networks*, CUP, 2009.

[Dempster, 1967] A. P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[Denœux, 2019] T. Denœux. Logistic regression, neural networks and Dempster–Shafer theory: A new perspective. *Knowledge-Based Systems*, 176:54–67, 2019.

[Fierens *et al.*, 2015] D. Fierens, G. den Broeck, J. Renkens, D. Shterionov, B. Gutmann, I. Thon, G. Janssens, and L. De Raedt. Inference and learning in probabilistic logic programs using weighted Boolean formulas. *TPLP*, 15(03):358–401, 2015.

[Gal and Ghahramani, 2016] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016.

[Gast and Roth, 2018] J. Gast and S. Roth. Lightweight probabilistic deep networks. In *CVPR*, pages 3369–3378, 2018.

[Haussmann *et al.*, 2021] M. Haussmann, S. Gerwinn, and M. Kandemir. Bayesian Evidential Deep Learning with PAC Regularization. In *AABI*, 2021.

[Hora, 1996] S. C. Hora. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety*, 54(2):217–223, 1996.

[Hougen *et al.*, 2021] C. D. Hougen, L. M. Kaplan, F. Cerutti, and A. O. Hero. Uncertain Bayesian Networks: Learning from Incomplete Data. In *MLSP*, 2021.

[Hüllermeier and Waegeman, 2021] E. Hüllermeier and W. Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Mach. Learn.*, 110(3):457–506, 2021.

[Ichikawa and Steup, 2018] J. J. Ichikawa and M. Steup. The Analysis of Knowledge. In *The Stanford Encyclopedia of Philosophy*. Summer, 2018.

[Jøsang, 2016] A. Jøsang. *Subjective Logic: A Formalism for Reasoning under Uncertainty*. Springer, 2016.

[Kaplan and Ivanovska, 2018] L. M. Kaplan and M. Ivanovska. Efficient belief propagation in second-order

Bayesian networks for singly-connected graphs. *IJAR*, 93:132–152, 2018.

[Kaplan *et al.*, 2020] L. M. Kaplan, F. Cerutti, M. Şensoy, and K. V. Mishra. Second-order learning and inference using incomplete data for uncertain Bayesian networks: A two node example. In *FUSION*, pages 1–8, 2020.

[Kim and Pearl, 1983] J. Kim and J. Pearl. A computational model for causal and diagnostic reasoning in inference systems. In *IJCAI*, 1983.

[Kimmig *et al.*, 2011] A. Kimmig, G. Van den Broeck, and L. De Raedt. An algebraic prolog for reasoning about possible worlds. In *AAAI*, pages 209–214, 2011.

[Kimmig *et al.*, 2017] A. Kimmig, G. Van den Broeck, and L. De Raedt. Algebraic model counting. *Journal of Applied Logic*, 22:46–62, 2017.

[Kingma and Welling, 2014] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

[Kisa *et al.*, 2014] D. Kisa, G. Van den Broeck, A. Choi, and A. Darwiche. Probabilistic Sentential Decision Diagrams. In *KR*, pages 558–567, 2014.

[Kocielnik *et al.*, 2019] R. Kocielnik, S. Amershi, and P. N. Bennett. Will you accept an imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *CHI*, 2019.

[Kopetzki *et al.*, 2021] A.-K. Kopetzki, B. Charpentier, D. Zügner, S. Giri, and S. Günnemann. Evaluating robustness of predictive uncertainty estimation: Are Dirichlet-based models reliable? In *ICML*, pages 5707–5718, 2021.

[Laplace, 1825] P. S. Laplace. *A Philosophical Essay on Probabilities*. Springer, 1825.

[Lauritzen, 1995] S. L. Lauritzen. The EM algorithm for graphical association models with missing data. *Computational Statistics & Data Analysis*, 19(2):191–201, 1995.

[Lecun *et al.*, 1998] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. of the IEEE*, 86(11):2278–2324, 1998.

[Louizos and Welling, 2017] C. Louizos and M. Welling. Multiplicative Normalizing Flows for Variational Bayesian Neural Networks. In *ICML*, pages 2218–2227, 2017.

[Ly *et al.*, 2017] A. Ly, M. Marsman, J. Verhagen, R. P. P. P. Grasman, and E.-J. Wagenmakers. A Tutorial on Fisher information. *J. Math. Psychol.*, 80:40–55, 2017.

[Mackay, 1995] D. J. C. Mackay. Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.

[Malinin and Gales, 2018] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *NIPS*, pages 7047–7058, 2018.

[Malinin and Gales, 2019] A. Malinin and M. Gales. Reverse KL-Divergence Training of Prior Networks: Improved Uncertainty and Adversarial Robustness. In *NeurIPS*, 2019.

[Martens, 2020] J. Martens. New Insights and Perspectives on the Natural Gradient Method. *JMLR*, 21(146):1–76, 2020.

[Mastrandrea *et al.*, 2011] M. D. Mastrandrea, K. J. Mach, G.-K. Plattner, O. Edenhofer, T. F. Stocker, C. B. Field, K. L. Ebi, and P. R. Matschoss. The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. *Climatic Change*, 108(4):675, 2011.

[McAllester, 2003] D. A. McAllester. PAC-Bayesian Stochastic Model Selection. *Mach. Learn.*, 51(1):5–21, 2003.

[Murphy, 2012] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[Peharz *et al.*, 2020] R. Peharz, A. Vergari, K. Stelzner, A. Molina, X. Shao, M. Trapp, K. Kersting, and Z. Ghahramani. Random Sum-Product Networks: A Simple and Effective Approach to Probabilistic Deep Learning. In *UAI*, pages 334–344, 2020.

[Poon and Domingos, 2011] H. Poon and P. Domingos. Sum-Product Networks: A New Deep Architecture. In *UAI*, pages 337–346, 2011.

[Rashwan *et al.*, 2016] A. Rashwan, H. Zhao, and P. Poupart. Online and Distributed Bayesian Moment Matching for Parameter Learning in Sum-Product Networks. In *AISTATS*, pages 1469–1477, 2016.

[Rezende and Mohamed, 2015] D. Jimenez Rezende and S. Mohamed. Variational inference with normalizing flows. In *ICML*, pages 1530–1538, 2015.

[Rohmer, 2020] J. Rohmer. Uncertainties in conditional probability tables of discrete Bayesian Belief Networks: A comprehensive review. *Eng. Appl. Artif. Intell.*, 88:103384, 2020.

[Şensoy *et al.*, 2018] M. Şensoy, L. M. Kaplan, and M. Kandemir. Evidential Deep Learning to Quantify Classification Uncertainty. In *NeurIPS*, 2018.

[Şensoy *et al.*, 2020] M. Şensoy, L. M. Kaplan, F. Cerutti, and M. Saleki. Uncertainty-Aware Deep Classifiers using Generative Models. In *AAAI*, pages 5620–5627, 2020.

[Shafer, 1976] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[Smets, 1993] P. Smets. Belief functions: The disjunctive rule of combination and the generalized Bayesian theorem. *IJAR*, 9(1):1–35, 1993.

[Valiant, 1979a] L. G. Valiant. The Complexity of Enumeration and Reliability Problems. *SIAM Journal on Computing*, 8(3):410–421, 1979.

[Valiant, 1979b] L. G. Valiant. The complexity of computing the permanent. *TCS*, 8(2):189–201, 1979.

[van der Vaart, 1998] A. W. van der Vaart. *Asymptotic Statistics*. CUP, 1998.

[Virani *et al.*, 2020] N. Virani, n. Iyer, and Z. Yang. Justification-Based Reliability in Machine Learning. *AAAI*, pages 6078–6085, 2020.