Asymptotic Analysis of Uncertain Naïve Bayes via Second-Order Probabilities

Lance M. Kaplan* and James Z. Hare*
* DEVCOM Army Research Lab, USA. {lance.m.kaplan.civ,james.z.hare.civ}@army.mil

***Abstract*—Likelihood fusion is a special case of Bayesian networks known as *naïve Bayes*. It is well known that as the number of observations goes to infinity with *known* likelihoods, the aleatoric uncertainty of the queried (or parent) variable goes to zero, and furthermore, the declared values are guaranteed to match the ground truth. This work considers the case that the conditional probabilities are learned with limited training data leading to *uncertain* likelihoods, and second-order probabilistic reasoning is incorporated to characterize the aleatoric and epistemic uncertainty. Remarkably, it is shown that both the aleatoric and epistemic uncertainty goes to zero despite limited knowledge of the likelihoods. The rate of convergence is dictated by a quasi-divergence value that is related to the Kullback-Liebler (KL) divergence. However, the quasi-divergence can be negative leading to false declarations. This paper investigates when false declarations can emerge and shows how such cases diminish as the amount of training data for the likelihoods increases.**

***Index Terms*—Uncertain Bayesian Networks, Second-Order Probabilities, Aleatoric and epistemic uncertainty**

I. Introduction

Bayes rule is the predominate basis of many information fusion processes [1] such as target tracking and object classification. While classification can entail the processing of sensor streams, it can also incorporate other contextual features such as the kinematic states of the unknown target. The NATO STANAG 4162 elucidates a wide array of contextual features that can be incorporated in an identification process [2], and at the heart of the process is fusion of likelihoods, which is simply the implementation of Bayes rule.

Fundamentally, likelihood fusion can be represented as a special case of a Bayesian network known as *naïve Bayes* (see Figure 1), which reasons probabilistically. Over the years, naïve Bayes and probabilistic approaches in general have been criticized for not being able to handle imprecision in the likelihoods (or conditional probabilities of the evidence variables) and conflicts between likelihoods [3]. In this light, imprecise versions of naïve Bayes have emerged, e.g., based on fuzzy sets [4], credal sets [5], and rough sets [6]. A survey of inference and learning methods for the more general imprecise Bayesian network problem is provided in [7].

Some researchers have argued that probabilistic Bayesian reasoning can in fact incorporate imprecision and conflict through proper modeling of the situations [8]. This work considers second-order probabilities [9] to develop *uncertain naïve Bayes* that models the imprecision of the conditional probabilities by the posterior distribution of these probabilities given a limited size training dataset. Recent work with general uncertain Bayesian networks have demonstrated empirically that such second-order probabilities lead to a better calibration of the desired confidence to capture the ground truth than credal and evidential theory approaches [10], [11].

This paper characterizes the performance of second-order probabilistic reasoning as the number of observational nodes providing measurements (see Figure 1) grows asymptotically. For precise naïve Bayes, the classic result is that asymptotically, the predicted value of the parent node will be confident and correct [12]. The theoretical results in this paper show that for uncertain naïve Bayes, predictions become both precise and confident asymptotically even in the presence of small training datasets. However, for small datasets, there are cases where the prediction is almost surely wrong.

The analysis in this paper is part of a larger effort of the Uncertainty Representation Working Group (ETURWG) to enhance the uncertainty representation and reasoning evaluation framework (URREF) with an understanding of the tradeoffs of various uncertainty representations [13]. The URREF helps information fusion system developers to properly incorporate uncertainty in their designs. Future work needs to consider the asymptotic performance of other representation for uncertain naïve Bayes fusion.

This paper is organized as follows. Section II reviews the precise naïve Bayes formulation, and Section III introduces the uncertain formulation. The theoretical results are presented in Section IV followed by examples that validate the theory in Section V. Section VI provides a discussion of the results, and finally, Section VII concludes the paper.

II. Precise Naïve Bayesian Networks

A Bayesian network is a directed graphical model that describes the joint probability mass function (pmf) of a set of random variables. The model then enables the determination of the posterior probability of target variables conditioned on the values of other observed variable values. Many sensor fusion problems can be modeled by a special Bayesian network known as *naïve Bayes*, which models the classification problem. Here, the target variable H is the parent to $\tilde{Z}$ variables whose values are observed (see Figure 1). The process of determining the posterior probability of the value of H given the values of the children nodes is known as likelihood fusion. The value of H can represent the target class, and the value of the children nodes enable an inference method to determine the posterior probabilities of the true target class.

In this work, we consider a binary valued parent H whose value $h \in \{0, 1\}$ is unknown and needs to be determined. The

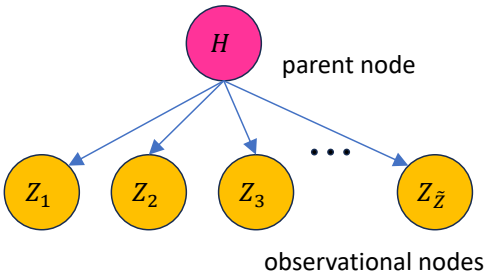


Figure 1: Structure of the naïve Bayesian network.

observation nodes are K-valued categorical variables such that the value of the i-th node, i.e., $z_i \in \{1, \cdots, K\}$, are drawn by a categorical distribution whose probabilities are conditioned on the value of the parent, i.e,

$$z_i \sim P_C(z|h) = \prod_{k=1}^{K} \left(p_{i,h,k}\right)^{\delta_{z,k}}, \tag{1}$$

where $p_{i,h,k} = p_{Z_i|H}(z_i = k|h)$ form the parameters of the Bayesian network and represents the probability that the observation from the i-th node takes the value k conditioned on the parent value being h. Note that $\delta_{z,k}$ is the Kronecker delta function that is zero unless $z = k$ in which case the value is one. Because the observations are drawn from different distributions if $h = 0$ or $h = 1$, it is possible to determine a posterior probability for h.

Overall, the joint pmf of the Bayesian network is

$$p(\mathbf{z}, h) = p_H(h) \prod_{i=1}^{\tilde{Z}} p_{i,h,z_i}, \tag{2}$$

where for ease of notation[1] $p_H(h = 0) = p_H(h = 1) = 0.5$.

For the classical Bayesian inference, the probability of the parent (or target) value given by the evidence is

$$p(h = 1|\mathbf{z}) = \frac{\prod_{i=1}^{\tilde{Z}} p_{i,1,z_i}}{\prod_{i=1}^{\tilde{Z}} p_{i,1,z_i} + \prod_{i=1}^{\tilde{Z}} p_{i,0,z_i}}. \tag{3}$$

It is well known [12] that as long as the conditional probabilities are correct, as the number of observations $\tilde{Z} \to \infty$, then likelihood fusion determines the true value of h with zero error, i.e., $p(h = 1|\mathbf{z}) = 0$ when $h = 0$ and $p(h = 1|\mathbf{z}) = 1$ when $h = 1$.

III. Uncertain Naive Bayesian Network

In practice, the conditional probabilities are not known *a priori* and must be learned from training data comprising of observations drawn from the i-th node when the value of parent h is known to be zero or one. Given enough training data, one can use the maximum likelihood estimates of the conditional probabilities as the true conditional parameters in the Bayesian network. However, with limited training data, it is crucial to consider the conditionals as random variables as answers to many queries may not be supported by the training data. As a result, second-order probabilities have been used to extend inference methods for Bayesian networks to determine a distribution for the queried probabilities [10], [11], [14].

The training data consists of the values of the variable values of all nodes in the Bayesian network over various independent instantiations of sampling from the joint pmf. Then, the posterior distribution of the conditional probabilities is determined by counting the instances of various children variable values when the parent variable takes on particular values. To determine the posterior distribution, let's consider a non-informative prior that the probabilities $\mathbf{p}$ for the K different values is uniformly distributed[2] over the simplex of possible values, i.e.,

$$\mathcal{S} = \left\{ \mathbf{p} \in \mathbb{R}^K_{\geq 0} : \sum_{k=1}^{K} p_k = 1 \right\}. \tag{4}$$

Given N instantiations where the parent variable takes value h and the counts for the occurrence of the various K values of the child is $\mathbf{n}$, which is in the discrete simplex of non-negative integers summing to N, i.e.,

$$\mathcal{N}_N = \left\{ \mathbf{n} \in \mathbb{Z}^K_{\geq 0} : \sum_{k=1}^{K} n_k = N \right\}, \tag{5}$$

then it is easy to show that the posterior distribution for the probabilities over the simplex $\mathcal{S}$ is Dirichlet distributed via

$$f_\beta(\mathbf{p}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \tag{6}$$

where the Dirichlet parameters $\alpha_k = n_k + 1$ for $k = 1, \ldots, K$, and $B(\boldsymbol{\alpha})$ is the K-dimensional beta function [15].

The Dirichlet distribution is a second-order probability of the categorical pmf. Any second-order probability distribution of the conditional probabilities encodes both aleatoric and epistemic uncertainty about the uncertain Bayesian network. Aleatoric uncertainty is the intrinsic randomness in the model and is based on the actual probability values, where the Shannon entropy of the probabilities is a reasonable quantification of aleatoric uncertainty. Note that aleatoric uncertainty is irreducible with respect to the training data as it is related to the ground truth probabilities. On the other hand, epistemic uncertainty represents the knowledge gap when learning the conditional probabilities from finite training data. While there are various ways to quantify these uncertainties, many researchers in the machine learning community computes the aleatoric (AU) and epistemic (EU) uncertainties as [16], [17]

$$\text{AU} = -\int_{\mathcal{S}} \sum_{k=1}^{K} p_k \log(p_k) f(\mathbf{p}) d\mathbf{p}, \tag{7a}$$

$$\text{EU} = -\sum_{k=1}^{K} \int_{\mathcal{S}} p_k f(\mathbf{p}) d\mathbf{p} \log\left(\int_{\mathcal{S}} p_k f(\mathbf{p}) d\mathbf{p} \right) - \text{AU}, \tag{7b}$$

[1] The asymptotic results in Theorem 1 hold for any non-extreme prior such that $p_H(h = 0) \in (0, 1)$.

[2] Other priors could be considered leading to less elegant analytical results.

where $f(\mathbf{p})$ is a second-order density of possible pmfs that need not be Dirichlet.

To train the uncertain naïve Bayes network, $N_{i,h}$ samples of instantiations of the value of node i are collected when the parent value is h. These instantiations are summarized as a histogram of possible values $\mathbf{n}_{i,h} \in \mathcal{N}_{N_{i,h}}$. These values are samples of the multinomial distribution whose parameters are the true conditional probabilities $\mathbf{p}_{i,h}$ so that

$$\mathbf{n}_{i,h} \sim P_M(\mathbf{n}; \mathbf{p}_{i,h}) = \binom{N}{n_1 \cdots n_k} \prod_{k=1}^{K} p_{i,h,k}^{n_k}. \quad (8)$$

Once training obtains the counts $\mathbf{n}_{i,h}$, the conditional probabilities are known within Dirichlet distributions (given by (6)) with parameters $\boldsymbol{\alpha}_{i,h} = \mathbf{n}_{i,h} + \mathbf{1}$.

Given the distributional knowledge of the conditional probabilities, second-order probabilities determine the distributions of the queried pmf via the non-linear relation given by (3), where the conditionals $\mathbf{p}_{i,h}$ are replaced with the corresponding random variables $\mathbf{P}_{i,h}$ that are distributed via (6). This means that the queried output pmf is actual a random variable $P(h = 1|\mathbf{z})$, but due to the nonlinear operation, the exact distribution is intractable. Prior work has investigated the δ-method as a means to fit the output distribution to a Dirichlet by propagating second-order statistics via linearization of the operations that transforms the conditionals to the output queried probabilities [10], [11], [14]. While the δ-method has proven to provide a well-calibrated expression of uncertainties about the queried outputs in empirical studies of small sized Bayesian networks [11], preliminary studies with the naïve Bayes network indicate that the accuracy of the δ-method drops as the number of children nodes $\tilde{Z}$ gets large.

The Monte Carlo method is a more accurate means to approximate the target query distribution than the δ-method, but it is more computationally complex. As such, it was used for gold standard comparisons in [11]. In short, Monte Carlo methods generates S instantiations of the conditional probabilities sampled from their respective Dirichlet posteriors. Then a standard first-order Bayesian inference method [18]–[20] determines the queried probabilities (e.g. (3)) for each instantiations. Finally, the first two moments (mean and variance) are computed for the output probabilities over the S instantiations and the Dirichlet distribution that best matches these moments approximates the true second-order distribution for the queried output.

For second-order inference of the uncertain naïve Bayes network, the s-th instantiation for the i-th node when the parent value is h is samples from the Dirichlet distribution as

$$\hat{\mathbf{p}}_{i,h}^{(s)} \sim f_\beta(\mathbf{p}; \mathbf{n}_{i,h} + \mathbf{1}). \quad (9)$$

The s-th queried probability $\hat{p}^{(s)}(h = 1|\mathbf{z})$ are computed via (3), where the conditional probabilities $p_{i,h,k}$ are replaced with the samples $\hat{p}_{i,h,k}^{(s)}$. The final stage approximates a Dirichlet distribution via moment matching using the collection of the S queried values.

The theory in the following section will consider the Monte Carlo approach to analyze the performance of the uncertain naïve Bayes network as the number of children $\tilde{Z} \to \infty$. Remarkably, all traces (or instantiations) almost surely lead to declaring the same confident decision (either $P(h = 1|\mathbf{z})$ is one or zero) meaning that both the declared aleatoric and epistemic uncertainties (as computed by (7a) and (7b)) are zero in par with the case of precisely known conditional probabilities. However, there are cases when the certain declaration is wrong. These cases become rarer as the amount of training data to learn the conditional probabilities gets larger.

IV. Theory

To show that both aleatoric and epistemic uncertainty goes to zero for uncertain naïve Bayes network queries, this section considers the convergence of the log-likelihood ratio for the two hypothesized values of the parent for an arbitrary Monte Carlo instantiation (trace) of conditional probabilities to determine one possible query result. Without loss of generality, we set the parent value to $h = 1$ so that $p(h = 1|\mathbf{z}) = 1$ is the correct declaration. It can be seen that the uncertain computation $p(h = 1|\mathbf{z})$ has a one-to-one correspondence with the log-likelihood by rearranging terms in (3) and replacing the conditional probabilities with their random variables s.t.

$$p^{(s)}(h = 1|\mathbf{z}) = \frac{1}{1 + e^{-\mathcal{L}^{(s)}(\mathbf{z})}}, \quad (10)$$

where $\mathcal{L}^{(s)}(\mathbf{z})$ is the s-th Monte Carlo instantiation of the sampled log-likelihood

$$\mathcal{L}^{(s)}(\mathbf{z}) = \sum_{i=1}^{\tilde{Z}} \ell_i^{(s)} = \sum_{i=1}^{\tilde{Z}} \log\left(\frac{\hat{p}_{i,1,z_i}^{(s)}}{\hat{p}_{i,0,z_i}^{(s)}}\right). \quad (11)$$

The log-likelihood $\ell_1^{(s)}$ due to the observation at the i-th node and the s-th trace is the result of the local observation z_i and the sampled conditional probabilities $\hat{\mathbf{p}}_{i,h}^{(s)}$. The observation z_i is a sampling via (1) once where $h = 1$. Each set of conditional probabilities $\hat{\mathbf{p}}_{i,0}^{(s)}$ and $\hat{\mathbf{p}}_{i,1}^{(s)}$ are sampled for each trace from the posterior Dirichlet distribution in (9) with corresponding trained counts of $\mathbf{n}_{i,0}$ and $\mathbf{n}_{i,1}$, respectively. Note that the training process to determine these counts are done once by effectively sampling the counts via (8) for $h = 0$ and $h = 1$. Overall, the generative process for $\ell_1^{(s)}$ is conditioned on the true conditional probabilities $\mathbf{p}_{i,0}$ and $\mathbf{p}_{i,1}$.

Essentially, the naïve Bayes is a likelihood test whose performance is driven by the expected log-likelihood at each node

$$D(\mathbf{p}_{i,1}, \mathbf{p}_{i,0}) = E\left[\log\left(\frac{\hat{p}_{i,1,z}}{\hat{p}_{i,0,z}}\right)\right]. \quad (12)$$

We refer to this expected log-likelihood as the *uncertain quasi-divergence* between the conditionals at the i-th node due to its analytical form as expressed by the following lemma. Note for ease of notation, we drop the explicit index for the node i in the remainder of the paper unless necessary for context.

Lemma 1. *Given that a node observation is sampled from a K-value multinomial with pmfs $\mathbf{p}_1$ and $\mathbf{p}_0$ and trained using N_1 and N_0 samples when the parent node has value $h = 1$ and $h = 0$, respectively, the expected log-likelihood when the true parent value $h = 1$ is*

$$D(\mathbf{p}_1, \mathbf{p}_0) = \sum_{k=1}^{K} p_{1,k} \left(L_{N_1}(p_{1,k}) - L_{N_0}(p_{0,k}) \right) + \sum_{i=1}^{K-1} \left(\frac{1}{N_0 + i} - \frac{1}{N_1 + i} \right), \tag{13}$$

where

$$L_N(p) = -\sum_{n=1}^{N} \frac{1}{n} (1-p)^n \tag{14}$$

is the N-th order power series approximation of the logarithm function.

Proof. See Appendix A. □

It is clear that as the training set sizes grow, $\lim_{N\to\infty} L_N(p) \to \log(p)$, and the second summation term in (13) goes to zero. Thus it is easy to see that the uncertain quasi-divergence asymptotically converges to the Kullback-Leibler divergence $D_{KL}(\mathbf{p}_1||\mathbf{p}_0)$. However, $D(\mathbf{p}_1, \mathbf{p}_0)$ is not a divergence in the strict sense as it can be negative, and there are cases when $D(\mathbf{p}_1, \mathbf{p}_0) = 0$ even though $\mathbf{p}_1 \neq \mathbf{p}_0$.

For using the uncertainty quasi-divergence for analysis, we need the following lemma.

Lemma 2. *Given a node observation as described in Lemma 1, the variance of the log-likelihood is bounded by a finite value, i.e.,*

$$VAR[\ell_i] < 2\psi^{(1)}(1) + (1 + \log(\max\{N_1, N_0\} + K))^2 < \infty. \tag{15}$$

Proof. See Appendix B. □

Note that $\psi^{(1)}(\cdot)$ is the trigamma function, which is the first derivative of $\psi(\cdot)$, and $\psi^{(1)}(1) \approx 1.6449$.

To enable analysis for possibly heterogeneous observational nodes in the network, we assume that all the observational nodes are K-valued multinomial distributed where their conditional pmfs are determined as independent identically distributed samples

$$\mathbf{p}_{i,1} \sim f_1(\mathbf{p}), \qquad \mathbf{p}_{i,2} \sim f_0(\mathbf{p}) \tag{16}$$

for all $i \geq 1$. As a result, the average uncertain quasi-divergence as the number of observational nodes grows asymptotically converges to a finite mean value, i.e.,

$$\lim_{\tilde{Z}\to\infty} \frac{1}{\tilde{Z}} \sum_{i=1}^{\tilde{Z}} D(\mathbf{p}_{i,1}, \mathbf{p}_{i,0}) \to \bar{D}, \tag{17}$$

where

$$\bar{D} = \int_{\mathcal{S}\times\mathcal{S}} D(\mathbf{p}_1, \mathbf{p}_0) f_1(\mathbf{p}_1) f_0(\mathbf{p}_0) d\mathbf{p}_1 d\mathbf{p}_0. \tag{18}$$

The generation of the log-likelihood value for Monte Carlo analysis at each of the observation nodes may appear to be identical: 1) sample the conditional pmfs, 2) sample the training counts, 3) sample the conditional pmf estimates. However, the values of $\ell_i^{(s)}$ for $i = 1, \ldots, \tilde{Z}$ are not identically distributed because the amount of training data can vary for each observational node. Nevertheless, we assume

$$\sup_i \max\{N_{i,1}, N_{i,0}\} = \tilde{N} < \infty. \tag{19}$$

The following theorem is the main result of this paper relating the asymptotic performance of queried parent probability values.

Theorem 1. *Asymptotically as $\tilde{Z} \to \infty$ the epistemic and aleatoric uncertainty of the naïve Bayes query result of (3) goes to zero declaring that $h = 1$ or $h = 0$. When the parent value $h = 1$, then the declaration is true when $\bar{D} > 0$, but it is false when $\bar{D} < 0$.*

Proof. Given that $\ell_i^{(s)}$ are independently distributed with variances bounded by $2\psi^{(1)}(1) + (1 + \log(\tilde{N} + K))^2 < \infty$, Kolmogorov's strong law of large numbers [21] shows that with probability 1,

$$\lim_{\tilde{Z}\to\infty} \frac{1}{\tilde{Z}} \mathcal{L}^{(s)} \to \bar{D}.$$

If $\bar{D} > 0$, then for any $\epsilon > 0$, there exist a value $\tilde{Z}_\epsilon$ such that

$$\frac{1}{\tilde{Z}} \mathcal{L}^{(s)} > \bar{D} - \epsilon > 0 \quad \Rightarrow \quad \mathcal{L}^{(s)} > \tilde{Z}(\bar{D} - \epsilon),$$

for all $\tilde{Z} \geq \tilde{Z}_\epsilon$ almost surely. Thus $\mathcal{L}^{(s)} \to \infty$ and $p^{(s)}(h = 1|\mathbf{z}) \to 1$ almost surely. Then for every Monte Carlo instantiation $p^{(s)}(h = 1|\mathbf{z}) = 1$ leading to a correct declaration with no epistemic and aleatoric uncertainty. Similarly, for $\bar{D} < 0$, then almost surely $\mathcal{L}^{(s)}$ goes to $-\infty$ at a rate of $\bar{D}$ leading to $p^{(s)}(h = 1|\mathbf{z}) = 0$ for every Monte Carlo instantiation, i.e., an incorrect declaration with no epistemic or aleatoric uncertainty □

The theorem shows that in almost all cases where the standard naïve Bayes asymptotically leads to a correct declaration with zero aleatoric uncertainty (epistemic uncertainty is always zero), the uncertain naïve Bayes also asymptotically achieves zero epistemic and aleatoric uncertainty. When $\mathbf{p}_{i,1} = \mathbf{p}_{i,0}$ and $N_1 = N_0$, the observations are uninformative and both D_{KL} and uncertain quasi-divergence are zero, leading to no refinement of uncertainty. Note, that there are additional cases leading to a zero uncertain quasi-divergence (see the boundaries in Figures 2(b) and (d)).

Unlike the classic result, the theorem also indicates that when the average quasi-divergence is negative, then the certain declaration is wrong. As the training data sizes $N_{i,h}$ grows, uncertain quasi-divergence gets closer to the KL divergence,

and the number of cases where it is negative decreases. For the condition that there is balance between the training data for the $h = 1$ and $h = 0$ cases, then the second term in (13) is zero. If there is more training data for the ground truth parent value $h = 1$ so that $N_{i,1} > N_{i,0}$, then the second term in (13) leads to a positive bias increasing the uncertain quasi-divergence and decreasing the cases where wrong declarations can occur. On the hand, if the training data is larger for the alternative parent value $h = 0$, then $N_{i,0} > N_{i,1}$, which leads to a negative bias in the uncertain quasi-divergence. In other word, training data mismatches leads to biasing the declaration in favor of the majority class, and when the true class is the minority class, this leads to a decrease of performance. The theorem provides a theoretical underpinning for the imbalanced training data problem [22], which will be illustrated in the next section.

V. Examples

For the sake of simplicity, this section analyzes the case that the conditional pmfs for each observational node are the same so that $f_1(\mathbf{p}) = \delta(\mathbf{p} - \mathbf{p}_1)$ and $f_0(\mathbf{p}) = \delta(\mathbf{p} - \mathbf{p}_0)$, respectively, where $\delta(\cdot)$ is the Dirac delta function. The general assumption is that the decision maker has no knowledge of the distribution f_1 and f_0 to leverage these distributions. If the decision maker knows that the pmfs are the same for each node, then one could trivially reduce epistemic uncertainty to zero by pooling together the training data. However, if the pmfs are not the same, this strategy could lead to bad results.

Figure 2(a) and (c) provides heatmaps for the uncertain quasi-divergence when the child nodes values are drawn from a binomial distributions for balance training data of size $N = 2$ and $N = 5$, respectively. The y- and x-axes represent the $k = 1$ value of the pmf for $h = 1$ and $h = 0$ values of the parent, respectively. Since the parent value is $h = 1$, the observations are being driven by the $\mathbf{p}_1$ values. Figure 2(b) and (d) show the regions in blue where the quasi-divergence is negative. The uncertain quasi-divergence is zero along the diagonal where $\mathbf{p}_1 = \mathbf{p}_0$. Slightly off the diagonal the quasi-divergence will dip below zero because the gradient is non-zero at the diagonal. The valley of negative values is deeper for the smaller training data size of $N = 2$. On the other hand, the divergence values rise faster for the larger $N = 5$ case. .

The regions of negative quasi-divergence increase when the training data is imbalanced in favor of the alternative ($h = 0$) distributions as shown in Figure 3. The figures show the cases when $N_0 = 100$, but the ground truth observed distributions are learned with only $N_1 = 2$ or $N_1 = 5$ samples. For the most part, the imbalance lowers the divergence except when $\mathbf{p}_0$ and $\mathbf{p}_1$ are extremely different. The region of negative divergence increases forming a convex region surrounding the diagonal case where the precise naïve Bayes formulation would be indeterminate but the imprecision and the training data imbalance leads to declaring the minority class. As the number of training data samples increases to $N = 5$, the region of negative divergence does recede.

When the training data imbalance is in favor of the true class, the bias increases the uncertain quasi-divergence. When

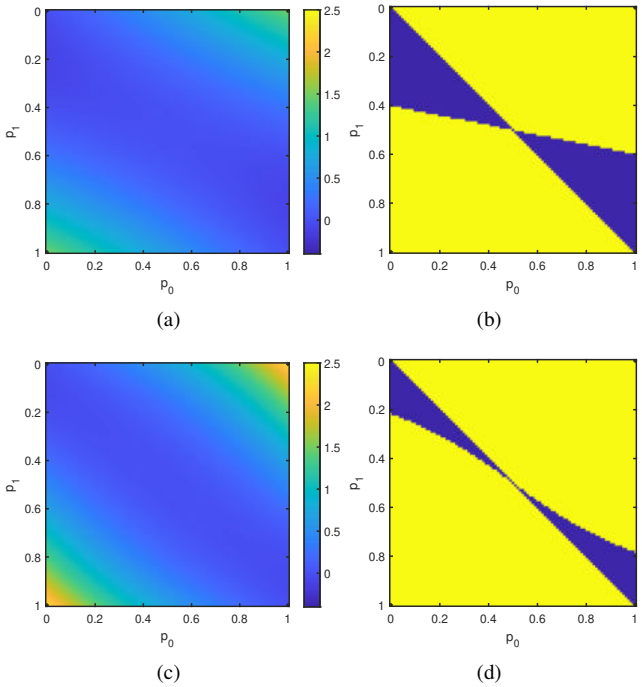


Figure 2: The uncertainty quasi-divergence with balanced training data for various values of the actual p_0 and alternative p_1 observation probabilities for (a) $N_1 = N_0 = 2$, (b) $N_1 = N_0 = 2$ (binarized), (c) $N_1 = N_0 = 2$, and (d) $N_1 = N_0 = 2$ (binarized).

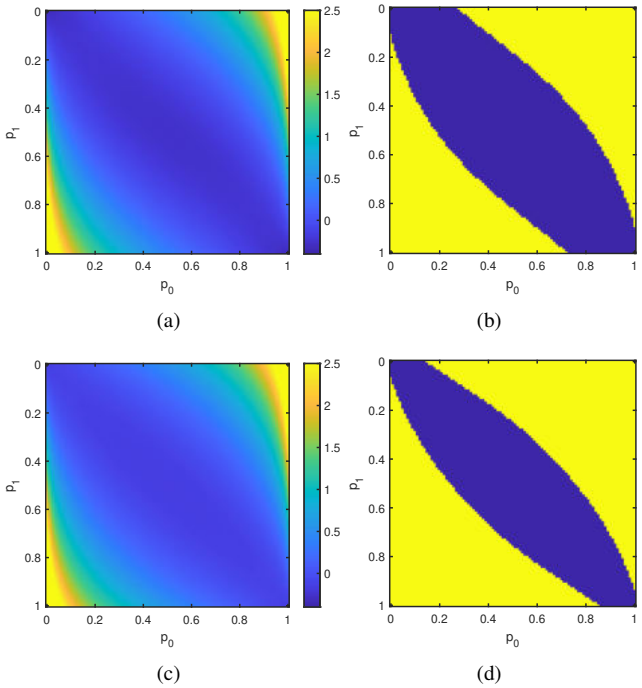
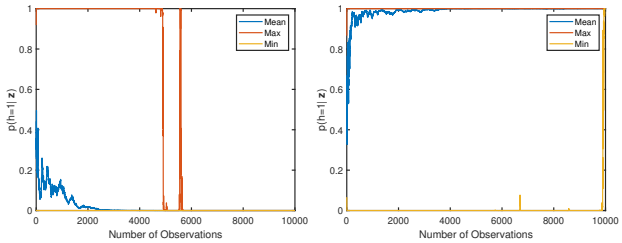
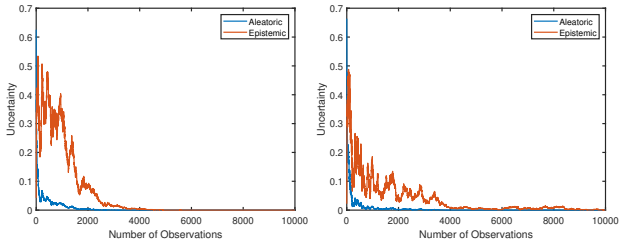


Figure 3: The uncertainty quasi-divergence with imbalanced data training for various values of the actual p_0 and alternative p_1 observation probabilities where $N_0 = 100$ for (a) $N_1 = 2$, (b) $N_1 = 2$ (binarized), (c) $N_1 = 5$, and (d) $N_1 = 5$ (binarized).



(a) (b)

Figure 4: The evolution of the value of the query probability $p(h = 1|\mathbf{z})$ via second-order Monte Carlo inference: (a) $N_1 = N_0 = 2$, and (b) $N_1 = N_0 = 5$.



(a) (b)

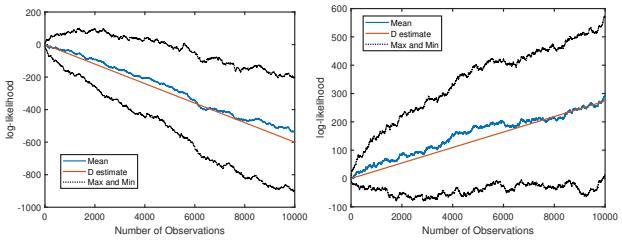
Figure 5: The evolution of the aleatoric and epistemic uncertainty for the query probability via second-order Monte Carlo inference: (a) $N_1 = N_0 = 2$, and (b) $N_1 = N_0 = 5$.

$N_1 = 100$ samples but the alternative class is sampled with $N_0 = 2$ or $N_0 = 5$ samples, the divergence shifts up except where $\mathbf{p}_0$ and $\mathbf{p}_1$ are extremely different. In both cases, the upward shift is enough for the negative divergence region to completely disappear. Due to space limitations, the plots are not provided.

To demonstrate that the uncertain quasi-divergence value dictates when inference of the uncertain Naïve Bayes leads to the correct declaration or not, we generated 10,000 traces of Monte Carlo inference of an uncertain network where the pmfs for the conditionals are all trained with $N = 2$ samples with $\mathbf{p}_1 = [.3, .7]^T$ and $\mathbf{p}_0 = [.1, .9]^T$. As seen in Figure 2(b), this condition is in the negative region with $D = -0.0600$. Figure 4(a) shows how the traces of the query output $p^{(s)}(h = 1|\mathbf{z})$ cumulatively evolves as more observations are included in the inference. The figure plots out the average trace along with the minimum and maximum values. All the probability traces are going to zero, which is the wrong declaration. Initially, there are traces confidently declaring either $h = 1$ or $h = 0$. After incorporating 6,000 observations all traces are making the wrong declaration as negative quasi-divergence predicts.

Figure 5 plots the evolution of the aleatoric and epistemic uncertainty as computed via (7a) and (7b), respectively. The plot indicates that around 2000 observations, all traces are confident (no aleatoric uncertainty), and by around 4,500 observations, the epistemic uncertainty is near zero, with only a handful of traces correctly declaring $h = 1$.

Figure 6 plots the evolution of the log-likelihood value as



(a) (b)

Figure 6: The evolution of the log-likelihood $\mathcal{L}$ for the query via second-order Monte Carlo inference: (a) $N_1 = N_0 = 2$, and (b) $N_1 = N_0 = 5$.

computed via (11). Specifically, the plots includes the average, minimum, and maximum value over all traces. Clearly, the log-likelihood values for all traces are trending down. The proof of $Theorem$ 1 indicates that the uncertain quasi-divergence approximates the slope of the log-likelihood, i.e., $\mathcal{L}^{(s)} \approx D\tilde{Z}$. The figure includes this approximation and shows that it does track with the average likelihood trace.

Now, let's consider the same case of $\mathbf{p}_1 = [.3, .7]^T$ and $\mathbf{p}_0 = [.1, .9]^T$ with a balanced training set, but where $N = 5$. Now according to Figure 2(d), this condition is in the positive region and the quasi-divergence $D = 0.0273$. Figures 4(b) indicates that the average probability increases correctly to a value of one by about 5,000 observations. Because the magnitude of the quasi-divergence is smaller, it takes almost 10,000 observations for all traces to be confidently correct. Figure 5(b) indicates that aleatoric and epistemic uncertainty are decreasing to zero. The epistemic decrease is slower than the $N = 2$ case, Finally, Figure 6(b) confirms that the log-likelihood is trending up at the rate predicted by the uncertain quasi-divergence.

VI. Discussion

Our prior work had indicated that second-order probabilities generated distributions for the queried probabilities that provides calibrated confidence bounds through empirical evaluations [10], [11]. The analysis in this paper indicates that this calibration breaks down asymptotically for the pathological cases when the second-order probability are precise (zero epistemic uncertainty) and confident (zero aleatoric uncertainty) but completely wrong. This means that the second-order probabilistic inference does not necessarily work with the inclusion of more children. This is a limitation as it means that tabulation of second-order probabilities are not always sufficient to properly capture uncertainty.

One may need to keep track of the training data sizes along with the number of observational nodes to determine if a declaration is uncertain. Alternatively, our recent work on the *Uncertain Likelihood Ratio* (ULR) may be able to to determine implicitly that there is lack of evidence from the training data to rule out the lower probability class [23]. Future work will consider various methods to determine when uncertain inferences can actually be confident or not.

VII. Conclusion

This paper analyzes the the asymptotic performance of second-order probabilistic inference of *uncertain naïve Bayes* networks. It turns out that asymptotically the aleatoric and the epistemic uncertainty go to zero. The decay rate of the uncertainty is dictated by the expected log-likelihood that forms the uncertain quasi-divergence. In most cases, especially as the training dataset size increases, the quasi-divergence is positive leading to correct inference. However, when the quasi-divergence is negative, the inference is wrong. Future work will expand the analysis to consider multiple classes, higher dimensionality $K > 2$, and continuous-valued observations.

Appendix A
Proof of Lemma 1

The expected value of the log-likelihood of the i-th node is

$$D(\mathbf{p}_1, \mathbf{p}_0) = \sum_{z,\mathbf{n}_0,\mathbf{n}_1} \phi(\mathbf{n}_1, \mathbf{n}_0) P_C(z; \mathbf{p}_1) \cdot \cdot P_M(\mathbf{n}_0; \mathbf{p}_0) P_M(\mathbf{n}_1; \mathbf{p}_1), \tag{20}$$

where $\phi(\mathbf{n}_1, \mathbf{n}_0)$ is the expected log-likelihood over the distribution of conditional probabilities, i.e.,

$$\begin{aligned}\phi(\mathbf{n}_1, \mathbf{n}_0) &= \mathop{E}_{\hat{\mathbf{p}}_1,\hat{\mathbf{p}}_0}\left[\sum_{k=1}^{K} \delta_{z,k}\left(\log\left(\frac{\hat{p}_{1,k}}{\hat{p}_{0,k}}\right)\right)\right] \\ &= \sum_{k=1}^{K} \delta_{z,k}\left((\psi(n_{1,k}+1) - \psi(N_1+K)) - (\psi(n_{0,k}+1) - \psi(N_0+K))\right),\end{aligned}$$

since $\mathop{E}_{\hat{\mathbf{p}}}[(\log(\hat{p}_k))] = \psi(n_k+1) - \psi(N+K)$. Note that $\psi(a) = \frac{d}{da}\log\Gamma(a)$ is the digamma function.

Taking the expectation of ϕ over the observation and counts $\mathbf{n}$ leads to

$$D(\mathbf{p}_1, \mathbf{p}_0) = \sum_{k=1}^{K} \mathop{E}_{z}[\delta_{z,k}] \mathop{E}_{\mathbf{n}_1}[\psi(n_{1,k}) - \psi(N_1+K)] \cdot \mathop{E}_{\mathbf{n}_0}[\psi(n_{0,k}) - \psi(N_0+K)]. \tag{21}$$

Clearly, the expectation of the categorical distribution $E[\delta_{z,k}] = p_{1,k}$. The expectation of the digammas over the multinomial can be derived for a generic pmf with pararameters $\mathbf{p}$ so that

$$\begin{aligned}&\mathop{E}_{\mathbf{n}}[\psi(n_z+1) - \psi(N+K)] \\ &= \sum_{\mathbf{n}\in\mathcal{N}_N} (\psi(n_z+1) - \psi(N+K)) P_M(\mathbf{n}; \mathbf{p}), \\ &= -\psi(N+K) + \sum_{n_z=0}^{K} \psi(n_z+1)\binom{N}{n_z} p_z^{n_z} \cdot \\ &\qquad \cdot \sum_{\mathbf{n}_{-z}\in\mathcal{N}_{N-n_z}} \frac{(N-n_z)!}{\prod_{i\neq z}(n_i)!} \prod_{k\neq z} p_k^{n_k}, \\ &= -\psi(N+K) + g(p_z),\end{aligned} \tag{22}$$

where

$$g(p_z) = \sum_{n_z=0}^{N} \psi(n_z+1)\binom{N}{n_z} p_z^{n_z}(1-p_z)^{N-n_z}.$$

To determine a simple closed form of $g(p_z)$, we take its first derivative to obtain

$$\begin{aligned}g'(p_z) &= \sum_{n_z=0}^{N} \psi(n_z+1)\binom{N}{n_z}\left(\frac{n_z}{p_z} - \frac{N-n_z}{1-p_z}\right) \cdot \\ &\qquad \cdot p_z^{n_z}(1-p_z)^{N-n_z}, \\ &= N\sum_{n_z=1}^{N} \psi(n_z+1)\binom{N-1}{n_z-1} p_z^{n_z-1}(1-p_z)^{N-n_z} \\ &\quad -N\sum_{n_z=0}^{N-1} \psi(n_z+1)\binom{N-1}{n_z} p_z^{n_z}(1-p_z)^{N-1-n_z}, \\ &= N\sum_{n_z=0}^{N-1} (\psi(n_z+2) - \psi(n_z+1)) \cdot \\ &\qquad \cdot \binom{N-1}{n_z} p_z^{n_z}(1-p_z)^{N-1-n_z}, \\ &= N\sum_{n_z=0}^{N-1} \frac{1}{n_z+1}\binom{N-1}{n_z} p_z^{n_z}(1-p_z)^{N-1-n_z}, \\ &= \frac{1}{p_z}\sum_{n_z=1}^{N} \binom{N}{n_z} p_z^{n_z}(1-p_z)^{N-n_z}, \\ &= \frac{1-(1-p_z)^N}{p_z} = \sum_{n=0}^{N-1}(1-p_z)^n.\end{aligned}$$

Integrating g' and noting that $g(1) = \psi(N+1)$ leads to

$$g(p_z) = \psi(N+1) + L_N(p_z), \tag{24}$$

where

$$L_N(p_z) = -\sum_{n=1}^{N} \frac{1}{n}(1-p_z)^n$$

is the N-th power series approximation of the logarithm function.

Plugging (24) into (22) leads

$$\mathop{E}_{\mathbf{n}}[\psi(n_z+1) - \psi(N+K)] = \psi(N+1) - \psi(N+K) + L_N(p_z).$$

Note that it is known that [24, Section 1.3]

$$\psi(N+1) - \psi(N+K) = -\sum_{i=1}^{K-1} \frac{1}{N+i}. \tag{25}$$

Now plugging in the expected values in (21) leads to the expression of $D(\mathbf{p}_1, \mathbf{p}_0)$ in (13). □

Appendix B
Proof of Lemma 2

The total variance of the log-likelihood can be broken down into parts via

$$\begin{aligned}\text{VAR}[\ell] &= E[(\ell - \phi(\mathbf{n}_1, \mathbf{n}_0) + \phi(\mathbf{n}_1, \mathbf{n}_0) - D(\mathbf{p}_1, \mathbf{p}_0))^2] \\ &= E[(\ell - \phi(\mathbf{n}_1, \mathbf{n}_0))^2] + E[(\phi(\mathbf{n}_1, \mathbf{n}_0) - D(\mathbf{p}_1, \mathbf{p}_0))^2] \\ &= E[\underset{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_0}{\text{VAR}}[\ell|\mathbf{n}_1, \mathbf{n}_0]] + \text{VAR}[\phi(\mathbf{n}_1, \mathbf{n}_0)]\end{aligned}$$

First, lets bound $\underset{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_0}{\text{VAR}}[\ell|\mathbf{n}_1, \mathbf{n}_0]$ uniformly with respect to $\mathbf{n}_1$ and $\mathbf{n}_0$ by noting that

$$\begin{aligned}\ell - \phi(\mathbf{n}_1, \mathbf{n}_0) = \\ \sum_{k=1}^{K} \delta_{z,k} \left(\log \hat{p}_{1,k} - \psi(n_{1,k}+1) + \psi(N_1 + K)\right) \\ - \delta_{z,k} \left(\log \hat{p}_{0,k} - \psi(n_{0,k}+1) + \psi(N_0 + K)\right),\end{aligned}$$

so that

$$\begin{aligned}\underset{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_0}{\text{VAR}}[\ell|\mathbf{n}_1, \mathbf{n}_0] &= \sum_{k=1}^{K} (\underset{\hat{\mathbf{p}}_1}{\text{VAR}}[\log \hat{p}_{1,k}] + \underset{\hat{\mathbf{p}}_0}{\text{VAR}}[\log \hat{p}_{0,k}]) \\ &= \sum_{k=1}^{K} \delta_{z,k} \Big(\psi^{(1)}(n_{1,k}+1) - \psi^{(1)}(N_1 + K) \\ &\quad + \psi^{(1)}(n_{0,k}+1) - \psi^{(1)}(N_0 + K)\Big) \\ &\leq \sum_{k=1}^{K} \delta_{z,k}(2\psi^{(1)}(1)) \leq 2\psi^{(1)}(1), \qquad (26)\end{aligned}$$

The above derivation is leveraging the relationship between the variance of the log of Dirichlet distributed probability values and the trigamma function, and the fact that the trigamma function is non-negative and monotonically decreasing for $\mathbb{R}_{\geq 0}$ [25].

Next, $\text{VAR}[\phi(\mathbf{n}_1, \mathbf{n}_0)]$ is bounded by simply using the upper bound for $\phi(\mathbf{n}_1, \mathbf{n}_0)$ in (21). Using the difference of digammas identify in (25), it is easy to see that for $0 \leq n \leq N$, the lower bound is

$$-(1+\log(N+K)) \leq -\sum_{i=1}^{N+K-1} \frac{1}{i} \leq \psi(n+1) - \psi(N+K),$$

and the upper bound is

$$\psi(n+1) - \psi(N+K) \leq -\sum_{i=N+1}^{N+K-1} \frac{1}{i} \leq 0.$$

This leads to the uniform upper bound

$$\phi^2(\mathbf{n}_1, \mathbf{n}_0) \leq (1 + \log(\max\{N_1, N_0\} + K))^2, \qquad (27)$$

which is also an upper bound for $\text{VAR}[\phi(\mathbf{n}_1, \mathbf{n}_0)]$. Overall the bounds (26) and (27) in relation to the total variance leads to the bound in (15). □

References

[1] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.

[2] E. Blasch, C. Yang, and I. Kadar, "Summary of tracking and identification methods," in *Signal Processing, Sensor/Information Fusion, and Target Recognition XXIII*, vol. 9091. SPIE, 2014, pp. 15–26.

[3] P. Walley, *Statistical reasoning with imprecise probabilities*. Springer, 1991, vol. 42.

[4] G. Y. Tütüncü and N. Kayaalp, "An aggregated fuzzy naive Bayes data classifier," *Journal of computational and applied mathematics*, vol. 286, pp. 17–27, 2015.

[5] M. Zaffalon, "The naive credal classifier," *Journal of statistical planning and inference*, vol. 105, no. 1, pp. 5–21, 2002.

[6] Y. Yao and B. Zhou, "Naive Bayesian rough sets," in *Rough Set and Knowledge Technology: 5th International Conference, RSKT 2010, Beijing, China, October 15-17, 2010. Proceedings 5*. Springer, 2010, pp. 719–726.

[7] J. Rohmer, "Uncertainties in conditional probability tables of discrete Bayesian belief networks: A comprehensive review," *Engineering Applications of Artificial Intelligence*, vol. 88, p. 103384, 2020.

[8] D. D. Mauá, D. Conaty, F. G. Cozman, K. Poppenhaeger, and C. P. de Campos, "Robustifying sum-product networks," *International Journal of Approximate Reasoning*, vol. 101, pp. 163–180, 2018.

[9] J. Baron, "Second-order probabilities and belief functions," *Theory and Decision*, vol. 23, pp. 25–36, 1987.

[10] L. Kaplan and M. Ivanovska, "Efficient belief propagation in second-order Bayesian networks for singly-connected graphs," *International Journal of Approximate Reasoning*, vol. 93, pp. 132–152, 2018.

[11] F. Cerutti, L. M. Kaplan, A. Kimmig, and M. Şensoy, "Handling epistemic and aleatory uncertainties in probabilistic circuits," *Machine Learning*, pp. 1–43, 2022.

[12] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[13] P. Costa, A.-L. Jousselme, K. B. Laskey, E. Blasch, V. Dragos, J. Ziegler, P. de Villiers, and G. Pavlin, "URREF: Uncertainty representation and reasoning evaluation framework for information fusion," *Journal of Advances in Information Fusion*, vol. 13, no. 2, pp. 137–157, 2018.

[14] T. Van Allen, A. Singh, R. Greiner, and P. Hooper, "Quantifying the uncertainty of a belief net response: Bayesian error-bars for belief net inference," *Artificial Intelligence*, vol. 172, no. 4-5, pp. 483–513, 2008.

[15] K. W. Ng, G.-L. Tian, and M.-L. Tang, *Dirichlet and related distributions: Theory, methods and applications*. John Wiley & Sons, 2011.

[16] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, "Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1184–1193.

[17] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.

[18] J. Pearl, "Fusion, propagation, and structuring in belief networks," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022, pp. 139–188.

[19] S. L. Lauritzen, "Propagation of probabilities, means, and variances in mixed graphical association models," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1098–1108, 1992.

[20] M. Chavira and A. Darwiche, "Compiling Bayesian networks using variable elimination." in *IJCAI*, vol. 2443. Citeseer, 2007.

[21] P. K. Sen and J. M. Singer, *Large sample methods in statistics (1994): An introduction with applications*. CRC press, 2017.

[22] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty based under-sampling for learning naive Bayes classifiers under imbalanced data sets," *IEEE Access*, vol. 8, pp. 2122–2133, 2019.

[23] J. Z. Hare and L. M. Kaplan, "Improved small sample hypothesis testing using the uncertain likelihood ratio," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[24] H. M. Srivastava and J. Choi, *Zeta and q-Zeta functions and associated series and integrals*. Elsevier, 2011.

[25] B. Schneider, "Algorithm as 121: trigamma function," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 27, no. 1, pp. 97–99, 1978.