On Bayesian Network Classifiers with Reduced Precision Parameters

Sebastian Tschiatschek, and Franz Pernkopf, *Senior Member, IEEE*

***Abstract*—Bayesian network classifiers (BNCs) are typically implemented on nowadays desktop computers. However, many real world applications require classifier implementation on embedded or low power systems. Aspects for this purpose have not been studied rigorously. We partly close this gap by analyzing reduced precision implementations of BNCs. In detail, we investigate the quantization of the parameters of BNCs with discrete valued nodes including the implications on the classification rate (CR). We derive worst-case and probabilistic bounds on the CR for different bit-widths. These bounds are evaluated on several benchmark datasets. Furthermore, we compare the classification performance and the robustness of BNCs with generatively and discriminatively optimized parameters, i.e. parameters optimized for high data likelihood and parameters optimized for classification, with respect to parameter quantization. Generatively optimized parameters are more robust for very low bit-widths, i.e. less classifications change because of quantization. However, classification performance is better for discriminatively optimized parameters for all but very low bit-widths. Additionally, we perform analysis for margin-optimized tree augmented network (TAN) structures which outperform generatively optimized TAN structures in terms of CR and robustness.**

***Index Terms*—Bayesian network classifiers, custom precision, quantization, discriminative learning**

I. Introduction

BAYESIAN network classifiers (BNCs) [3] are probabilistic classifiers. In many scenarios, they achieve classification rates (CRs) similar to state-of-the-art classifiers, e.g. support vector machines (SVMs) [4]. BNCs are widely used in different fields, e.g. in expert systems for the medical domain [5], for classifying gene expression data in microbiology [6], and in many other areas such as speech and image processing. An advantage of BNCs over other classifiers is a compact model representation — using less parameters than other classifiers while achieving comparable CR performance [4].

BNCs consist of a directed acyclic graph (DAG) whose nodes correspond to random variables (RVs), and a set of conditional probability densities (CPDs) associated with those nodes. The DAG comprises the structure of the BNC and encodes independences between the RVs. Assuming discrete valued RVs, the CPDs can be represented as conditional probability tables (CPTs). In this case, the values of these CPTs are the parameters of the BNC.

S. Tschiatschek and F. Pernkopf are with the Department of Electrical Engineering, Laboratory of Signal Processing and Speech Communication, Graz University of Technology, Austria.
E-mail: {tschiatschek,pernkopf}@tugraz.at

This work was supported by the Austrian Science Fund (project number P25244-N15).

This work extends results published in [1], [2].

Most commonly BNCs are implemented on nowadays desktop computers, where double precision floating-point numbers are used for parameter representation and arithmetic operations. In general, these computations are considered as exact. However, to support the usage of BNCs in everyday applications, e.g. acoustic environment classification in hearing aids, these classifiers must be implemented efficiently in embedded or low power systems. One key aspect for an efficient implementation is the parametrization of BNCs using reduced precision parameters, e.g. fixed-point parameters with limited precision. For example, low bit-width parameters enable one to implement many BNCs in parallel on field programmable gate arrays (FPGAs), i.e. the circuit area requirements on the FPGA correlate with the parameter precision [7].

Results from sensitivity analysis indicate that BNCs are well suited for low bit-widths implementations because they are not sensitive to parameter deviations under the following two conditions [8]. Firstly, if the conditional probabilities are not too extreme, i.e. close to zero or one, and, secondly, if the posterior probabilities for different classes are significantly different. Additionally, this is supported by our empirical classification results for BNCs with reduced precision parameters [9], [1]; for certain datasets ~ 10 bits for representing the parameters and performing classification are sufficient to achieve CRs close to that of BNCs with double precision parameters. Furthermore, considering precision constraints during parameter learning is advantageous and can result in BNCs with better CR performance [10].

In this paper, we present novel theoretical results and extended empirical results for BNCs with finite precision fixed-point parameters. All our results are based on the assumption that parameters are learned in full-precision and rounded to the desired precision for classification. We derive three types of bounds on the classification performance after parameter precision reduction and compare these in experiments. Additionally, we empirically compare the classification performance and robustness of BNCs with respect to precision reduction for different learning paradigms. In particular, we use generatively and discriminatively optimized parameters/structures [3], [11], [12], [13], [14], [4]; let C denote the class variable to which objects represented by a set of features $\mathbf{X}$ are to be assigned, and let $P(C, \mathbf{X})$ be a joint probability density over C and $\mathbf{X}$. Then, *generatively optimized* means that model parameters/structures are optimized for a good fit of the data, i.e. the likelihood $P(C, \mathbf{X})$ of the data is maximized; and for classification the class conditional distribution $P(C|\mathbf{X}) = \frac{P(C,\mathbf{X})}{P(\mathbf{X})}$ is used. In contrast, *discriminatively optimized* means that model parameters/structures are optimized for good classification, i.e.

the class conditional distribution $P(C|\mathbf{X})$ is (directly) maximized either by the maximum conditional likelihood (MCL) or maximum margin (MM) objective. For generative parameter learning, we resort to Bayesian parameter estimation [15], [16]. This type of parameter learning results in a posterior distribution over the parameters, enabling us to consider the uncertainty in the parameter estimates in our bounds. Taking this uncertainty into account is crucial as the common assumptions of uniform and independent quantization error are incorrect and result in loose bounds. However, maximum likelihood (ML) estimates do not provide this uncertainty information.

Our main results presented in this paper are:

- Derivation of probabilistic and worst-case bounds on the classification performance of BNCs with quantized parameters.
- An empirical evaluation of these bounds on classical machine learning datasets.
- Empiric evidence that BNCs with discriminatively optimized parameters are *not* more robust[1] to parameter quantization than BNCs with generatively optimized parameters. However, classification performance using discriminatively optimized parameters is better when using bit-widths of 3 bits or more.
- Empiric evidence that BNCs with generatively optimized parameters and discriminatively optimized structures, i.e. structures optimized using a large-margin score, 1) yield higher classification performance, and 2) are more robust to parameter quantization than BNCs with generatively optimized structures. The former statement holds for all considered bit-widths (1–10 bits), and the latter statement holds for small bit-widths (5/10 bits, depending on the dataset).

This paper is structured as follows: In Section II we introduce our notation, formally introduce BNCs, and describe methods for assessing the CPTs of BNCs. In Section III we derive bounds on the CR performance of BNCs with reduced precision parameters and present experiments supporting our arguments in Section IV. We conclude the paper in Section V.

II. BACKGROUND

A. Probabilistic Classification

In probabilistic classification one assumes an RV C denoting the class and RVs $X_1, \ldots, X_L$ representing the attributes of the classifier. These RVs are modeled by a joint probability distribution $\mathrm{P}^*(C, \mathbf{X})$, where $\mathbf{X} = [X_1, \ldots, X_L]$ is a random vector. Typically, $\mathrm{P}^*(C, \mathbf{X})$ is unknown. However, a training set $\mathcal{D}$ consisting of N samples drawn i.i.d. from $\mathrm{P}^*(C, \mathbf{X})$ is available, i.e. $\mathcal{D} = \{(c^{(n)}, \mathbf{x}^{(n)}) | n = 1, \ldots, N\}$, where $c^{(n)}$ denotes the instantiation of C and $\mathbf{x}^{(n)}$ the instantiation of $\mathbf{X}$ in the n^{th} training sample. The aim is to induce *good* classifiers given $\mathcal{D}$. Formally, a classifier $h\colon \mathrm{sp}(\mathbf{X}) \to \mathrm{sp}(C)$ is a mapping, where $\mathrm{sp}(\mathbf{X})$ denotes the set of all assignments of $\mathbf{X}$ and $\mathrm{sp}(C)$ is the set of classes. The CR of this classifier is

$$\mathrm{CR}(h) := \mathbb{E}_{\mathrm{P}^*(C,\mathbf{X})}\left[\mathbf{1}(C = h(\mathbf{X}))\right], \tag{1}$$

where $\mathbf{1}(A)$ denotes the indicator function and $\mathbb{E}_{\mathrm{P}^*(C,\mathbf{X})}[\cdot]$ is the expectation operator with respect to the distribution $\mathrm{P}^*(C, \mathbf{X})$. The indicator function $\mathbf{1}(A)$ equals one if statement A is true and zero otherwise. Typically, the CR can not be evaluated but is estimated using cross-validation [17].

Any probability distribution $\mathrm{P}(C, \mathbf{X})$ naturally induces a classifier $h_{\mathrm{P}(C,\mathbf{X})}$, given as

$$\begin{aligned} h_{\mathrm{P}(C,\mathbf{X})}\colon \mathrm{sp}(\mathbf{X}) &\to \mathrm{sp}(C), \\ \mathbf{x} &\mapsto \arg\max_{c \in C} \mathrm{P}(C = c | \mathbf{X} = \mathbf{x}). \end{aligned} \tag{2}$$

In this way, each instantiation $\mathbf{x}$ of $\mathbf{X}$ is classified as the maximum a-posteriori (MAP) estimate of C given $\mathbf{x}$ under $\mathrm{P}(C, \mathbf{X})$.

B. Learning Bayesian Network Classifiers

Bayesian networks (BNs) [18], [19] are used to represent joint probability distributions in a compact and intuitive way. A BN $\mathcal{B} = (\mathcal{G}, \mathcal{P}_\mathcal{G})$ consists of a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{V} = \{X_0, \ldots, X_L\}$ is the set of nodes and $\mathbf{E}$ the set of edges of the graph, and a set of local CPDs $\mathcal{P}_\mathcal{G} = \{\mathrm{P}(X_0|\Pi_{X_0}), \ldots, \mathrm{P}(X_L|\Pi_{X_L})\}$. The terms $\Pi_{X_0}, \ldots, \Pi_{X_L}$ denote the set of parents of $X_0, \ldots, X_L$ in $\mathcal{G}$, respectively. Each node of the graph corresponds to an RV and the edges of the graph determine dependencies between these RVs. Throughout this paper, we consider X_0 as the class variable. Whenever convenient, we denote X_0 as C. We assume that C has no parents in $\mathcal{G}$, i.e. $\Pi_C = \emptyset$. A BN represents a joint probability $\mathrm{P}^\mathcal{B}(C, X_1, \ldots, X_L)$ as product of local conditional distributions, i.e.

$$\mathrm{P}^\mathcal{B}(C, X_1, \ldots, X_L) = \mathrm{P}(C) \prod_{i=1}^{L} \mathrm{P}(X_i|\Pi_{X_i}). \tag{3}$$

In this paper, we assume discrete RVs $X_i \in \{1, \ldots, |\mathrm{sp}(X_i)|\}$.

BNs for classification [20], [3], i.e. BNCs, can be optimized in two ways: (i) By selecting the graph structure $\mathcal{G}$, i.e. structure learning [3], [21], [11]; (ii) By selecting the set of CPDs $\mathcal{P}_\mathcal{G}$, i.e. parameter learning [13], [14], [4]. Throughout this paper, we consider naive Bayes (NB) and tree augmented network (TAN) structures [3] only.

For learning the parameters $\mathcal{P}_\mathcal{G}$ of a BN two paradigms exist, namely generative parameter learning and discriminative parameter learning: In *generative parameter learning* one aims at identifying parameters representing the generative process leading to the data of the training set, i.e. the joint distribution $\mathrm{P}(C, \mathbf{X})$ is estimated from the training data. An example of this paradigm is ML learning. Its objective is maximization of the likelihood of the data with respect to the parameters. ML parameter learning is not well suited for our needs, as it only results in a point estimate of the parameters and does not capture information on the distribution of the parameters. However, this distribution is important when investigating effects of precision reduction. Therefore, we resort to Bayesian

[1]Robustness compares the number of changing classifications of generatively and discriminatively optimized BNCs with respect to parameter quantization.

parameter estimation using Dirichlet priors [16], [22]. We assume global parameter independence, i.e. the CPTs corresponding to the different nodes in the BNC are independent, and local parameter independence, i.e. the parameters for different parent states are independent [16]. In detail, assuming Dirichlet priors, the parameters[2] $\mathrm{P}(X_i|\Pi_{X_i}=\mathbf{h})$ follow a Dirichlet distribution $\mathrm{Dir}\left(\widetilde{\boldsymbol{\alpha}}^i_{\mathbf{h}}\right)$ with concentration parameters

$$\widetilde{\boldsymbol{\alpha}}^i_{\mathbf{h}} = \left[\widetilde{\alpha}^i_{1,\mathbf{h}},\ldots,\widetilde{\alpha}^i_{|\mathrm{sp}(X_i)||\mathbf{h}}\right], \tag{4}$$

i.e. $\mathrm{P}(X_i|\Pi_{X_i}=\mathbf{h}) \sim \mathrm{Dir}\left(\widetilde{\boldsymbol{\alpha}}^i_{\mathbf{h}}\right)$. Given the training data $\mathcal{D}$, the posterior parameter distribution is

$$\mathrm{P}(X_i|\Pi_{X_i}=\mathbf{h};\mathcal{D}) \sim \mathrm{Dir}\left(\widetilde{\boldsymbol{\alpha}}^i_{\mathbf{h}}+\mathbf{n}^i_{\mathbf{h}}\right), \tag{5}$$

where $\mathbf{n}^i_{\mathbf{h}}$ is a vector of frequency counts obtained from the training data. The j^{th} entry of $\mathbf{n}^i_{\mathbf{h}}$, denoted as $n^i_{j|\mathbf{h}}$, is the number of times $X_i=j$ together with $\Pi_{X_i}=\mathbf{h}$ is observed in the training data. Each parameter *instantiation* is marginally beta distributed, i.e.

$$\Theta^i_{j|\mathbf{h}} := \mathrm{P}(X_i=j|\Pi_{X_i}=\mathbf{h}) \sim \mathrm{Beta}\left(\alpha^i_{j|\mathbf{h}},\beta^i_{j|\mathbf{h}}\right), \tag{6}$$

where

$$\alpha^i_{j|\mathbf{h}} = \widetilde{\alpha}^i_{j|\mathbf{h}} + n^i_{j|\mathbf{h}}, \text{ and} \tag{7}$$

$$\beta^i_{j|\mathbf{h}} = \sum_{\substack{j'=1\\ j'\neq j}}^{|\mathrm{sp}(X_i)|} \left(\widetilde{\alpha}^i_{j'|\mathbf{h}} + n^i_{j'|\mathbf{h}}\right). \tag{8}$$

From these beta distributions, parameters with maximum a-posteriori probability (MAP parameters) can be computed as

$$\theta^{i,\mathrm{MAP}}_{j|\mathbf{h}} = \frac{\alpha^i_{j|\mathbf{h}}-1}{\alpha^i_{j|\mathbf{h}}+\beta^i_{j|\mathbf{h}}-2}. \tag{9}$$

In contrast to generative parameter learning, *discriminative parameter learning* aims at identifying parameters leading to good CR performance. Discriminative learning is especially useful if the distribution $\mathrm{P}^*(C,\mathbf{X})$ is not well represented by the considered BNs. To optimize the CR, the focus of learning is optimization of the conditional distribution $\mathrm{P}^{\mathcal{B}}(C|\mathbf{X})$ for classification. Several objectives for this purpose are known in the literature, e.g. the maximum conditional likelihood [14] objective and the MM [23], [4] objective. Throughout this paper, we consider the MM objective as a representative for discriminative parameter learning.

MM parameters $\mathcal{P}^{\mathrm{MM}}_{\mathcal{G}}$ are found as

$$\mathcal{P}^{\mathrm{MM}}_{\mathcal{G}} = \arg\max_{\mathcal{P}_{\mathcal{G}}} \prod_{n=1}^{N} \min\left(\gamma, d^{(n)}\right), \tag{10}$$

where $\min(\gamma,d^{(n)})$ denotes the hinge loss and $d^{(n)}$ is the margin of the n^{th} sample given as

$$d^{(n)} = \frac{\mathrm{P}^{\mathcal{B}}(c^{(n)},\mathbf{x}^{(n)})}{\max_{c\neq c^{(n)}} \mathrm{P}^{\mathcal{B}}(c,\mathbf{x}^{(n)})}, \tag{11}$$

[2]For convenience, we denote the RVs $\mathrm{P}(X_i|\Pi_{X_i}=\mathbf{h})$ as parameters.

and $\gamma>1$ is a parameter scaling the margin. In this way, the margin *measures* the likelihood of the n^{th} sample belonging to the correct class $c^{(n)}$ in relation to the strongest competing class. The n^{th} sample is correctly classified if $d^{(n)}>1$ and vice versa. This type of learning results in a point estimate for the parameters, i.e. no information on the distribution of the parameters is obtained.

III. Performance Bounds

In this section, we derive worst-case and probabilistic bounds on the classification rate of BNCs with CPTs represented by reduced precision fixed-point numbers. Before deriving these bounds, we formalize the considered scenario.

A. Setting

When representing the parameters of BNCs in reduced precision, one has to decide whether to represent the probabilities or the logarithmic probabilities. Typically, log probabilities are favored for numerical reasons, i.e. a large dynamic range is achieved and classification resorts to a simple addition. However, reduced precision log probabilities have the drawback that the resulting CPTs are in general not normalized. In contrast, when representing probabilities in reduced precision, ensuring proper normalization is easy. Note that to overcome this normalization issue, undirected graphical models are an appealing alternative as well [24]. As we only consider classification tasks in this paper, we focus on reduced precision log probabilities.

The logarithm of the joint probability induced by a BN $\mathcal{B}$ according to (3) is

$$\log \mathrm{P}^{\mathcal{B}}(\mathbf{X}=\mathbf{x}) = \sum_{i=0}^{L} \log \mathrm{P}(X_i=x_i|\Pi_{X_i}=\mathbf{x}_{\Pi_{X_i}}) \tag{12}$$

$$= \sum_{i=0}^{L}\sum_{\mathbf{h}}\sum_{j} \mathbf{1}(x_i=j,\mathbf{x}_{\Pi_{X_i}}=\mathbf{h}) W^i_{x_i|\mathbf{x}_{\Pi_{X_i}}=\mathbf{h}} \tag{13}$$

$$= \boldsymbol{\phi}(\mathbf{x})^T\mathbf{W}, \tag{14}$$

where $\boldsymbol{\phi}(\mathbf{x})$ collects the terms $\mathbf{1}(x_i=j,\mathbf{x}_{\Pi_{X_i}}=\mathbf{h})$ in a vector and $\mathbf{W}$ the terms $W^i_{x_i|\mathbf{x}_{\Pi_{X_i}}} := \log\Theta^i_{x_i|\mathbf{x}_{\Pi_{X_i}}}$, respectively; the expressions x_i and $\mathbf{x}_{\Pi_{X_i}}$ denote the instantiations of X_i and Π_{X_i} according to $\mathbf{x}$. Consequently, the probability $\mathrm{P}^{\mathcal{B}}(\mathbf{X}=\mathbf{x})$ can be written as

$$\mathrm{P}^{\mathcal{B}}(\mathbf{X}=\mathbf{x}) = e^{\boldsymbol{\phi}(\mathbf{x})^T\mathbf{W}}. \tag{15}$$

In the remainder of the paper, we consider the effect of quantizing the RVs $\mathbf{W}$ using fixed-point numbers with B bits. Quantization of these RVs is performed by rounding, i.e. let $w^i_{j|\mathbf{h}}$ denote an instantiation of $W^i_{j|\mathbf{h}}$ and $\mathcal{Q}^B_S(w^i_{j|\mathbf{h}})$ its quantized value using a scale factor $S>0$. Then

$$\mathcal{Q}^B_S(w^i_{j|\mathbf{h}}) = 2^{-B}\left[\frac{w^i_{j|\mathbf{h}}/S}{2^{-B}}\right]_R, \tag{16}$$

where $[a]_R$ means that a is rounded to the closest integer number. Note that we can arbitrarily scale all parameters by

some constant factor S without changing the class assignment of any sample. The usefulness and selection of the scale factor S is explained in the next section. In the remainder of this paper we denote the width of the quantization interval as q, i.e. $q = 2^{-B}$. Additionally, for ease of notation we write $\mathcal{Q}(\cdot)$ instead of $\mathcal{Q}_S^B(\cdot)$ and assume some fixed S and B.

B. Selection of the Scale Factor

By appropriately selecting the scale factor S, we ensure that all parameters can be represented using a desired number of B bits — without the need for integer bits. This alleviates our analysis since only the fractional bits of the fixed-point representation of all parameters have to be considered. Note that in general the scale factor should be as small as possible to ensure that the scaled parameters exploit the full range of representable numbers. For example, if we use a 5 bit fixed-point representation and the parameters are in the range $[0, -20]$, then these parameters should rather be scaled to $[0, -(1-2^{-5})]$ than to $[0, -0.5]$.

For generative or discriminative parameters, S is selected as follows:

a) Generative parameters: We use MAP scaled log parameters in the reduced precision BNCs. Let $Z \sim \mathrm{Beta}\big(\alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\big)$ with shape parameters $\alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}} > 1$. Further, let $Y = \frac{1}{S}\log(Z)$ be the corresponding scaled log parameters, where $S > 0$. The distribution $f_Y^S(y)$ of Y can be easily determined as

$$f_Y^S(y) = S\frac{e^{\alpha^i_{j|\mathbf{h}} Sy}(1-e^{Sy})^{\beta^i_{j|\mathbf{h}}-1}}{\mathrm{B}\big(\alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\big)}, \tag{17}$$

where $\mathrm{B}(c,d) = \int_0^1 u^{c-1}(1-u)^{d-1}\mathrm{d}u$ is the beta function. Assuming that we quantize Y using B bits and that quantization extends over the negative real numbers, the possible values are $\mathcal{C}_B = \{-n \cdot 2^{-B} \colon n \in \mathbb{N}_0\}$. For all $n \in \mathbb{N}_0$, the probability of $\mathcal{Q}(Y) = -nq$ is [25]

$$\mathrm{P}(\mathcal{Q}(Y) = -nq) = \int_{-nq-\frac{q}{2}}^{\min\{0, -nq+\frac{q}{2}\}} f_Y^S(y)\mathrm{d}y \tag{18}$$

$$= \frac{\mathrm{B}\Big(e^{S\min\{0,-nq+\frac{q}{2}\}}; \alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\Big) - \mathrm{B}\Big(e^{-Snq-S\frac{q}{2}}; \alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\Big)}{\mathrm{B}\big(\alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\big)}, \tag{19}$$

where $\mathrm{B}(k;c,d) = \int_0^k u^{c-1}(1-u)^{d-1}\mathrm{d}u$ is the incomplete beta function. The distribution $f_Y^S(y)$ is unimodal, i.e. its only maximum m is attained[3] at

$$m = \frac{1}{S}\log\frac{\alpha^i_{j|\mathbf{h}}}{\alpha^i_{j|\mathbf{h}} + \beta^i_{j|\mathbf{h}} - 1}. \tag{20}$$

Hence, we can easily determine the quantized parameter value with highest a-posteriori probability as

$$n^* = \arg\max_{\widetilde{n}\in\mathbb{N}_0} \mathrm{P}(\mathcal{Q}(Y) = -\widetilde{n}q) \tag{21}$$

$$= \arg\max_{n\in\{\lfloor m/q\rfloor, \lceil m/q\rceil\}} \mathrm{P}(\mathcal{Q}(Y) = nq), \tag{22}$$

[3]The maximum can be determined by solving $\frac{\mathrm{d}}{\mathrm{d}y} f_Y^S(y) = 0$.

where $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the floor and ceiling functions, respectively. To ensure that n^* can be represented using B bits, $S > 0$ is selected such that

$$|\arg\max_{n\in\{\lfloor m/q\rfloor, \lceil m/q\rceil\}} \mathrm{P}(\mathcal{Q}_S^B(W^i_{j|\mathbf{h}}) = nq)| \leq 2^B - 1. \tag{23}$$

Hence, we need that $|\lfloor m/q\rfloor| \leq 2^B - 1$ (note that m is negative). This inequality certainly holds if $-m/q \leq 2^B - 1$. Thus it suffices to select S such that

$$S \geq \max_{i,j,\mathbf{h}} \left[-\frac{1}{1-q}\log\left(\frac{\alpha^i_{j|\mathbf{h}}}{\alpha^i_{j|\mathbf{h}} + \beta^i_{j|\mathbf{h}} - 1}\right)\right]. \tag{24}$$

b) Discriminative parameters: Only a point estimate of the MM parameters $\mathbf{w}^{\mathrm{MM}}$ is available. We scale these parameters such that they can be represented using B bits. Hence, we require $|w^{i,\mathrm{MM}}_{j|\mathbf{h}}|/S \leq 1 - 2^{-B}$ for all $|w^{i,\mathrm{MM}}_{j|\mathbf{h}}|$. This is satisfied by selecting

$$S \geq \max_{i,j,\mathbf{h}} \frac{|w^{i,\mathrm{MM}}_{j|\mathbf{h}}|}{1-q}. \tag{25}$$

C. On the Quantization Error

Given the training data and having performed parameter estimation, we want to obtain bounds on the CR of BNCs when using fixed-point numbers with B bits to represent their CPTs. As explained in Section II, the components of the random vector $\mathbf{\Theta} = \exp(\mathbf{W})$ for Bayesian parameter estimates are marginally beta-distributed. A first observation is that in general the quantization error of the logarithmic parameters $\mathbf{W}$ is not uniform. This is especially true, when only few bits are used. For example, consider Figure 1; the distribution of the quantization error over quantization interval q of an RV $\log(Y)$, where Y is beta-distributed is shown. For convenience, the scale factor introduced above is set to $S = 1$. The RVs Y corresponding to the left and right subfigures are distributed as $Y \sim \mathrm{Beta}\big(1\cdot 10^3, 9\cdot 10^3\big)$ and $Y \sim \mathrm{Beta}\big(1\cdot 10^5, 9\cdot 10^5\big)$, respectively. When using sufficiently many bits, the quantization error is uniformly distributed in the quantization interval q. However, in the case of few bits, the distribution of the quantization error is not uniform. Consequently, assuming a uniformly distributed quantization error can be inappropriate. Furthermore, note that although both RVs have the same expected value, their quantization errors are different. As a consequence, also the average quantization error, indicated by the vertical red line in Figure 1, differs.

The setting considered above describes the situation when a fixed but unknown distribution of some RV is estimated from training sets of different sizes. The more training samples are observed, the more peaked the beta distribution becomes and the more concentrated is the quantization error (for small number of bits). This matches our intuition — the more training samples are available, the less uncertain are the parameter estimates. Consequently, the uncertainty about the expected quantization error is reduced with increasing sample size. This observation is reflected in the bounds derived below.

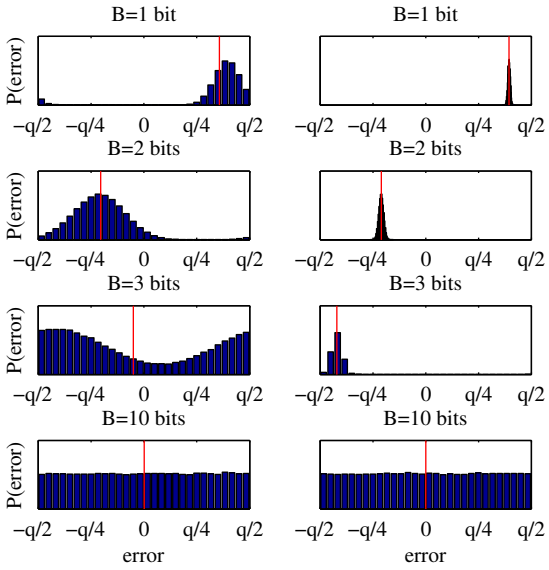


Fig. 1. Quantization error of $\log(Y)$, where Y is beta-distributed. Error histograms are computed from 10^5 samples. Left: $Y \sim \mathrm{Beta}\big(1 \cdot 10^3, 9 \cdot 10^3\big)$; Right: $Y \sim \mathrm{Beta}\big(1 \cdot 10^5, 9 \cdot 10^5\big)$; in both cases $\mathbb{E}_{\mathrm{P}(Y)}\left[\log Y\right] = -2.3026$; q is the quantization interval width; average quantization error (vertical red line)

D. Worst-Case Bound

We are now able to derive a worst-case bound on the CR. For ease of notation, we state this bound for the single-sample case:

Theorem 1 (Worst-Case Bound)**.** *Let $P^{\mathcal{B}}$ be the probability distribution defined by a BN, and let S, B, q be as introduced above. Further, let $(c, \mathbf{x})$ be a sample belonging to class c with feature instantiation $\mathbf{x}$. Assuming that the parameters of the BN are independent, the expected classification rate for this sample can be lower bounded as*

$$\mathbb{E}\left[\mathbf{1}(c = h_{P^{\mathcal{B}}}(\mathbf{x}))\right] \geq \tag{26}$$
$$1 - \min\left\{1, \sum_{c' \neq c} e^{S(L+1)q} \left[\prod_{(i,j,\mathbf{h}) \in A_1} M^i_{j|\mathbf{h}}\right]\left[\prod_{(i,j,\mathbf{h}) \in A_2} \widetilde{M}^i_{j|\mathbf{h}}\right]\right\},$$

where

$$A_1 = \{(i,j,\mathbf{h}) : \phi(c,\mathbf{x})^i_{j|\mathbf{h}} = 1\}, \tag{27}$$
$$A_2 = \{(i,j,\mathbf{h}) : \phi(c',\mathbf{x})^i_{j|\mathbf{h}} = 1\}, \tag{28}$$
$$M^i_{j|\mathbf{h}} = \frac{\alpha^i_{j|\mathbf{h}} + \beta^i_{j|\mathbf{h}} - 1}{\alpha^i_{j|\mathbf{h}} - 1}, \text{ and} \tag{29}$$
$$\widetilde{M}^i_{j|\mathbf{h}} = \frac{\alpha^i_{j|\mathbf{h}}}{\alpha^i_{j|\mathbf{h}} + \beta^i_{j|\mathbf{h}}}. \tag{30}$$

Before proceeding to the proof, we want to make some comments on the above bound: Typically this bound is not tight and rather conservative. Nevertheless, it allows for simple determination of a worst-case performance, by noting that the term

$$\left[\prod_{(i,j,\mathbf{h}) \in A_1} M^i_{j|\mathbf{h}}\right]\left[\prod_{(i,j,\mathbf{h}) \in A_2} \widetilde{M}^i_{j|\mathbf{h}}\right] \tag{31}$$

is constant for different choices of bits used for quantization. Hence, the lower bound in (26) can be evaluated by performing parameter learning once, computing the terms $M^i_{j|\mathbf{h}}$ and $\widetilde{M}^i_{j|\mathbf{h}}$ for all samples and classes and then performing a weighting by the exponential function $e^{S(L+1)q}$.

Proof of Theorem 1: The expected CR of a sample belonging to class c having feature instantiation $\mathbf{x}$ using BNC $\mathcal{B}$ with log parameters $\mathbf{W}$ can be lower bounded by

$$\mathbb{E}\left[\mathbf{1}(c = h_{P^{\mathcal{B}}}(\mathbf{x}))\right]$$
$$= \mathrm{P}\left(\boldsymbol{\phi}(c,\mathbf{x})^T \mathcal{Q}(\mathbf{W}) > \max_{c' \neq c} \boldsymbol{\phi}(c',\mathbf{x})^T \mathcal{Q}(\mathbf{W})\right) \tag{32}$$
$$= 1 - \mathrm{P}\left(\bigcup_{c' \neq c}\left[\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T \mathcal{Q}(\mathbf{W}) \leq 0\right]\right) \tag{33}$$
$$\overset{(a)}{\geq} 1 - \min\left\{1, \sum_{c' \neq c} \mathrm{P}\left(\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T \mathcal{Q}(\mathbf{W}) \leq 0\right)\right\} \tag{34}$$
$$\overset{(b)}{\geq} 1 - \min\left\{1, \sum_{c' \neq c} \mathrm{P}\left(\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T \frac{\mathbf{W}}{S} \leq q(L+1)\right)\right\}, \tag{35}$$

where the expectation is with respect to the distribution of the parameters $\mathbf{W}$. Inequality (a) is by the union bound and (b) is because

$$\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T \mathcal{Q}(\mathbf{W}) \leq 0 \tag{36}$$

implies

$$\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T \frac{\mathbf{W}}{S} \leq 2(L+1)\frac{q}{2} \tag{37}$$

assuming the *worst-case quantization error of* $\frac{q}{2}$ for each $\frac{W^i_{j|\mathbf{h}}}{S}$. We can further bound the above expression by determining Chernoff-type bounds on the terms of the form

$$\mathrm{P}\left(\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T \mathbf{W}/S \leq (L+1)q\right). \tag{38}$$

In a first step, for $t > 0$ we obtain

$$(38) = \mathrm{P}\left(e^{-t\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T \mathbf{W}} \geq e^{-t(L+1)qS}\right). \tag{39}$$

As $t > 0$ can be arbitrarily selected,

$$(38) \overset{(a)}{\leq} \inf_{t>0} e^{t(L+1)qS} \prod_{(i,j,\mathbf{h}) \in A_1} \mathbb{E}_{\mathrm{P}(W^i_{j|\mathbf{h}})}\left[e^{-tW^i_{j|\mathbf{h}}}\right] \cdot \prod_{(i,j,\mathbf{h}) \in A_2} \mathbb{E}_{\mathrm{P}(W^i_{j|\mathbf{h}})}\left[e^{tW^i_{j|\mathbf{h}}}\right] \tag{40}$$
$$\overset{(b)}{\leq} e^{S(L+1)q} \prod_{(i,j,\mathbf{h}) \in A_1} \mathbb{E}\left[e^{-W^i_{j|\mathbf{h}}}\right] \prod_{(i,j,\mathbf{h}) \in A_2} \mathbb{E}\left[e^{W^i_{j|\mathbf{h}}}\right] \tag{41}$$
$$= e^{S(L+1)q}\left[\prod_{(i,j,\mathbf{h}) \in A_1} M^i_{j|\mathbf{h}}\right]\left[\prod_{(i,j,\mathbf{h}) \in A_2} \widetilde{M}^i_{j|\mathbf{h}}\right], \tag{42}$$

where (a) is by Markov's inequality and independence of the parameters[4], and (b) is by arbitrarily selecting $t = 1$; $A_1 =$

[4]The assumed independence does not hold exactly for the RVs in $\mathbf{W}$ corresponding to the class node. Further, it does not hold for nodes independent of the class node. This later case however is uninteresting, as the introduced quantization error affects all classes in the same way.

$\{(i,j,\mathbf{h}) : \phi(c,\mathbf{x})^i_{j|\mathbf{h}} = 1\}$, $A_2 = \{(i,j,\mathbf{h}) : \phi(c',\mathbf{x})^i_{j|\mathbf{h}} = 1\}$. Further, $\widetilde{M}^i_{j|\mathbf{h}} = \frac{\alpha^i_{j|\mathbf{h}}}{\alpha^i_{j|\mathbf{h}}+\beta^i_{j|\mathbf{h}}}$, i.e. the expectation of the beta-distributed RV $e^{W^i_{j|\mathbf{h}}}$. Furthermore, $M^i_{j|\mathbf{h}}$ is computed as follows: The RV $e^{-W^i_{j|\mathbf{h}}}$ is distributed as the RV $Z := 1/Y$, where $Y \sim \mathrm{Beta}\left(\alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\right)$. The density $f_Z(z)$ can be computed [26] as

$$f_Z(z) = \begin{cases} 0 & 0 \le z \le 1, \\ \frac{(1/z)^{\alpha^i_{j|\mathbf{h}}+1}(1-1/z)^{\beta^i_{j|\mathbf{h}}-1}}{\mathrm{B}\left(\alpha^i_{j|\mathbf{h}},\beta^i_{j|\mathbf{h}}\right)} & 1 < z. \end{cases} \tag{43}$$

Hence, the expected value evaluates to

$$\begin{aligned} M^i_{j|\mathbf{h}} &= \mathbb{E}_{f_Z}[Z] && (44)\\ &= \int_1^\infty z f_Z(z)\mathrm{d}z && (45)\\ &= \frac{1}{\mathrm{B}\left(\alpha^i_{j|\mathbf{h}},\beta^i_{j|\mathbf{h}}\right)} \frac{\Gamma(\alpha^i_{j|\mathbf{h}}-1)\Gamma(\beta^i_{j|\mathbf{h}})}{\Gamma(\alpha^i_{j|\mathbf{h}}+\beta^i_{j|\mathbf{h}}-1)} && (46)\\ &= \frac{\alpha^i_{j|\mathbf{h}}+\beta^i_{j|\mathbf{h}}-1}{\alpha^i_{j|\mathbf{h}}-1}, && (47) \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function and where we used $\mathrm{B}(c,d) = \frac{\Gamma(c)\Gamma(d)}{\Gamma(c+d)}$ and the recurrence equation $\Gamma(c) = \frac{\Gamma(c+1)}{c}$.

The final bound is derived by using (42) in (35), i.e.

$$\mathbb{E}\left[\mathbf{1}(c = h_{P\mathcal{B}}(\mathbf{x}))\right] \ge 1 - \min\left\{1, \sum_{c'\neq c} e^{S(L+1)q}\left[\prod_{(i,j,\mathbf{h})\in A_1} M^i_{j|\mathbf{h}}\right]\left[\prod_{(i,j,\mathbf{h})\in A_2} \widetilde{M}^i_{j|\mathbf{h}}\right]\right\}. \tag{48}$$

■

E. Probabilistic Bound

The bound according to Theorem 1 is conservative and is, empirically, often very loose, cf. the experiments in Section IV. However, we can obtain a tighter bound by considering the stochastic nature of the quantization error. Due to quantization, each of the entries in $\mathbf{W}$ is distorted by a quantization error, i.e.

$$\mathcal{Q}(W^i_{j|\mathbf{h}}) = \frac{W^i_{j|\mathbf{h}}}{S} + E^i_{j|\mathbf{h}}, \tag{49}$$

where $E^i_{j|\mathbf{h}}$ is the error introduced by quantization. Note, that the error is a deterministic function of the scaled log parameters $\frac{1}{S}\mathbf{W}$. Therefore, it depends on both the number of bits B and the scale factor S. The error $E^i_{j|\mathbf{h}}$ can assume values in $[-\frac{q}{2}, +\frac{q}{2}]$, cf. Figure 1. Similar to before, the expected classification rate of a sample belonging to class c with feature instantiation $\mathbf{x}$ can be bounded:

Theorem 2 (Probabilistic Bound)**.** *Let $P^{\mathcal{B}}$ be the probability distribution defined by a BN, and let S, B, q be as introduced above. Further, let $(c,\mathbf{x})$ be a sample belonging to class c with feature instantiation $\mathbf{x}$. Assuming that the parameters of the BN are independent, the expected classification rate for this sample can be lower bounded as*

$$\mathbb{E}\left[\mathbf{1}(c = h_{P\mathcal{B}}(\mathbf{x}))\right] \ge 1 - \min\left\{1, \sum_{c'\neq c} F^{2(L+1)}\left[\prod_{(i,j,\mathbf{h})\in A_1} B^i_{j|\mathbf{h}}\right]\left[\prod_{(i,j,\mathbf{h})\in A_2} \widetilde{B}^i_{j|\mathbf{h}}\right]\right\}, \tag{50}$$

where $F = \frac{e^{Sq/2}-e^{-Sq/2}}{Sq}$, where

$$B^i_{j|\mathbf{h}} = \frac{1}{\mathrm{B}\left(\alpha^i_{j|\mathbf{h}},\beta^i_{j|\mathbf{h}}\right)} \sum_{k=0}^{\infty} e^{Skq} \cdot \left[\mathrm{B}\left(e^{S\min(0,-k\cdot q+q/2)}; \alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\right) - \mathrm{B}\left(e^{S(-kq-q/2)}; \alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\right)\right], \tag{51}$$

$$\widetilde{B}^i_{j|\mathbf{h}} = \frac{1}{\mathrm{B}\left(\alpha^i_{j|\mathbf{h}},\beta^i_{j|\mathbf{h}}\right)} \sum_{k=0}^{\infty} e^{-Skq} \cdot \left[\mathrm{B}\left(e^{S\min(0,-k\cdot q+q/2)}; \alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\right) - \mathrm{B}\left(e^{S(-kq-q/2)}; \alpha^i_{j|\mathbf{h}}, \beta^i_{j|\mathbf{h}}\right)\right], \tag{52}$$

and where A_1, A_2 are as in Theorem 1.

Before proceeding to the proof, we want to make a brief remark: For many combinations of α, β and q, the quantities $B^i_{j|\mathbf{h}}$ and $\widetilde{B}^i_{j|\mathbf{h}}$ can be approximated as

$$B^i_{j|\mathbf{h}} \approx \frac{1}{Sq} - \frac{e^{-Sq/2}}{Sq}\frac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)} - \frac{1}{Sq}\left[\frac{\mathrm{B}\left(e^{-Sq/2};\alpha,\beta\right)}{\mathrm{B}(\alpha,\beta)} - e^{Sq/2}\frac{\mathrm{B}\left(e^{-Sq/2};\alpha+1,\beta\right)}{\mathrm{B}(\alpha,\beta)}\right]. \tag{53}$$

and

$$\widetilde{B}^i_{j|\mathbf{h}} \approx -\frac{1}{Sq} + \frac{e^{-Sq/2}}{Sq}\frac{\mathrm{B}(\alpha-1,\beta)}{\mathrm{B}(\alpha,\beta)} - \frac{1}{Sq}\left[e^{-Sq/2}\frac{\mathrm{B}\left(e^{-Sq/2};\alpha-1,\beta\right)}{\mathrm{B}(\alpha,\beta)} - \frac{\mathrm{B}\left(e^{-Sq/2};\alpha,\beta\right)}{\mathrm{B}(\alpha,\beta)}\right]. \tag{54}$$

Details on these approximations are provided after the proof. In cases in which these approximations are not accurate, Equations (51) and (52) can be approximated by partial sums with only few terms.

Proof of Theorem 2: The expected classification rate of a sample belonging to class c with feature instantiation $\mathbf{x}$ is given as

$$\mathbb{E}\left[\mathbf{1}(c = h_{P\mathcal{B}}(\mathbf{x}))\right] = \mathrm{P}\left(\boldsymbol{\phi}(c,\mathbf{x})^T\mathcal{Q}(\mathbf{W}) > \max_{c'\neq c}\boldsymbol{\phi}(c,\mathbf{x})^T\mathcal{Q}(\mathbf{W})\right) \tag{55}$$

$$\ge 1 - \min\left\{1, \sum_{c'\neq c}\mathrm{P}\left(\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T\mathcal{Q}(\mathbf{W}) \le 0\right)\right\}, \tag{56}$$

where the quantization error is included in $\mathcal{Q}(\mathbf{W})$. Chernoff-type bounds for the terms

$$\mathrm{P}\left(\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T\mathcal{Q}(\mathbf{W}) \le 0\right) \tag{57}$$

read as

$$(57) \le \left[\prod_{(i,j,\mathbf{h})\in A_1}\mathbb{E}\left[e^{S\mathcal{Q}(W^i_{j|\mathbf{h}})}\right]\right]\left[\prod_{(i,j,\mathbf{h})\in A_2}\mathbb{E}\left[e^{-S\mathcal{Q}(W^i_{j|\mathbf{h}})}\right]\right] \tag{58}$$

For ease of notation, let $Y = \log(Z)$ and $Z \sim \mathrm{B}(\alpha, \beta)$. Then, the quantities in (58) read as

$$\mathbb{E}\left[e^{\pm S\mathcal{Q}(Y)}\right] = \int_{-\infty}^{0} e^{\pm Sq[y/q]_R} f_Y^S(y)\mathrm{d}y, \tag{59}$$

where Y corresponds to $W^i_{j|\mathbf{h}}$. As $e^{\pm Sq[y/q]_R}$ is constant for all y in any fixed quantization interval, we obtain

$$(59) = \sum_{k=-\infty}^{0} \int_{kq-q/2}^{\min(0,kq+q/2)} e^{\pm Skq} f_Y^S(y)\mathrm{d}y \tag{60}$$

$$\overset{(a)}{=} \sum_{k=0}^{\infty} \frac{e^{\mp Skq}}{\mathrm{B}(\alpha,\beta)} \int_{e^{S(-kq-q/2))}}^{e^{S\min(0,-kq+q/2)}} w^{\alpha-1}(1-w))^{\beta-1}\mathrm{d}w \tag{61}$$

$$= \frac{1}{\mathrm{B}(\alpha,\beta)} \sum_{k=0}^{\infty} e^{\mp Skq} \cdot \mathrm{B}(w;\alpha,\beta)\Big|_{e^{S(-kq-q/2)}}^{e^{S\min(0,-kq+q/2)}}, \tag{62}$$

where (a) is by substituting $w := e^{Sy}$. Hence, using (62) and (58) in (56) results in the desired bound. ∎

In the following, we present the derivation of the approximations for the terms $B^i_{j|\mathbf{h}}$ and $\widetilde{B}^i_{j|\mathbf{h}}$ mentioned above.

Approximations: Consider Equation (62). For the lower limit and the negative-sign case of the term $e^{\mp Skq}$,

$$\sum_{k=0}^{\infty} e^{-Skq}\mathrm{B}\Big(e^{S(-kq-q/2)};\alpha,\beta\Big) \tag{63}$$

$$\overset{(a)}{\approx} \int_0^{\infty} e^{-Sxq}\mathrm{B}\Big(e^{-Sqx}e^{-Sq/2};\alpha,\beta\Big)\,\mathrm{d}x \tag{64}$$

$$\overset{(b)}{=} \frac{1}{Sq}\int_0^1 \mathrm{B}\Big(ye^{-Sq/2};\alpha,\beta\Big)\,\mathrm{d}y \tag{65}$$

$$\overset{(c)}{=} \frac{1}{Sq}\left[\mathrm{B}\Big(e^{-Sq/2};\alpha,\beta\Big) - \int_0^1 (ye^{-Sq/2})^{\alpha}(1-ye^{-Sq/2})^{\beta-1}\mathrm{d}y\right] \tag{66}$$

$$\overset{(d)}{=} \frac{1}{Sq}\left[\mathrm{B}\Big(e^{-Sq/2};\alpha,\beta\Big) - e^{Sq/2}\int_0^{e^{-Sq/2}} z^{\alpha}(1-z)^{\beta-1}\mathrm{d}z\right] \tag{67}$$

$$= \frac{1}{Sq}\left[\mathrm{B}\Big(e^{-Sq/2};\alpha,\beta\Big) - e^{Sq/2}\mathrm{B}\Big(e^{-Sq/2};\alpha+1,\beta\Big)\right], \tag{68}$$

where (a) is by approximating the sum by an integral, (b) by substituting $y = e^{-Sqx}$, (c) by integration by parts, and (d) by substituting $z = ye^{-Sq/2}$. Thus,

$$\frac{1}{\mathrm{B}(\alpha,\beta)}\sum_{k=0}^{\infty} e^{-Skq}\mathrm{B}\Big(e^{S(-2k-1)q/2};\alpha,\beta\Big) \tag{69}$$

$$\approx \frac{1}{Sq}\left[\frac{\mathrm{B}\big(e^{-Sq/2};\alpha,\beta\big)}{\mathrm{B}(\alpha,\beta)} - e^{Sq/2}\frac{\mathrm{B}\big(e^{-Sq/2};\alpha+1,\beta\big)}{\mathrm{B}(\alpha,\beta)}\right]. \tag{70}$$

Similarly, for the upper limit in (62),

$$\frac{1}{\mathrm{B}(\alpha,\beta)}\sum_{k=0}^{\infty} e^{-Skq}\mathrm{B}\Big(\min(1, e^{S(-2k+1)q/2});\alpha,\beta\Big) \tag{71}$$

$$\approx \frac{1}{Sq} - \frac{e^{-Sq/2}}{Sq}\frac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)}. \tag{72}$$

Hence, using (68) and (72) in (62) we obtain

$$B^i_{j|\mathbf{h}} \approx \frac{1}{Sq} - \frac{e^{-Sq/2}}{Sq}\frac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)} - \frac{1}{Sq}\left[\frac{\mathrm{B}\big(e^{-Sq/2};\alpha,\beta\big)}{\mathrm{B}(\alpha,\beta)} - e^{Sq/2}\frac{\mathrm{B}\big(e^{-Sq/2};\alpha+1,\beta\big)}{\mathrm{B}(\alpha,\beta)}\right]. \tag{73}$$

Analogously, we obtain

$$\widetilde{B}^i_{j|\mathbf{h}} \approx -\frac{1}{Sq} + \frac{e^{-Sq/2}}{Sq}\frac{\mathrm{B}(\alpha-1,\beta)}{\mathrm{B}(\alpha,\beta)} - \frac{1}{Sq}\left[e^{-Sq/2}\frac{\mathrm{B}\big(e^{-Sq/2};\alpha-1,\beta\big)}{\mathrm{B}(\alpha,\beta)} - \frac{\mathrm{B}\big(e^{-Sq/2};\alpha,\beta\big)}{\mathrm{B}(\alpha,\beta)}\right]. \tag{74}$$

F. Probabilistic Bound Assuming Uniform and Independent Quantization Errors

In the following, to emphasize the need for considering an accurate model of the quantization error in (49), we determine CR bounds by assuming that the quantization error is uniform and independent of $\mathbf{W}$. This assumption is common although often inappropriate when analyzing quantization effects. In experiments, cf. Section IV, the resulting bounds are looser than the ones in Theorem 2. The bound can be stated as follows:

Theorem 3 (Uniform and Independent Error Bound)**.** *Let $P^{\mathcal{B}}$ be the probability distribution defined by a BN, and let S, B, q be as introduced. Further, let $(c, \mathbf{x})$ be a sample belonging to class c with feature instantiation $\mathbf{x}$. Assuming that the parameters of the BN are independent and that the quantization errors of the parameters are independent and uniformly distributed within a quantization interval, the expected classification rate for this sample can be lower bounded as*

$$\mathbb{E}\left[\mathbf{1}(c = h_{P^{\mathcal{B}}}(\mathbf{x}))\right] \geq 1 - \min\left\{1, \sum_{c'\neq c} F^{2(L+1)}\left[\prod_{(i,j,\mathbf{h})\in A_1} M^i_{j|\mathbf{h}}\right]\left[\prod_{(i,j,\mathbf{h})\in A_2} \widetilde{M}^i_{j|\mathbf{h}}\right]\right\}, \tag{75}$$

where $F = \frac{e^{Sq/2}-e^{-Sq/2}}{Sq}$, and A_1, A_2, $M^i_{j|\mathbf{h}}$, $\widetilde{M}^i_{j|\mathbf{h}}$ are as in Theorem 1.

Proof: Using B bits for quantization, $E^i_{j|\mathbf{h}} \sim U(\pm\frac{q}{2})$, where $U(\pm a)$ denotes a uniform distribution on the interval $[-a, +a]$. Hence,

$$\mathbb{E}_{E^i_{j|\mathbf{h}}\sim U(\pm\frac{q}{2})}\left[e^{SE^i_{j|\mathbf{h}}}\right] = \mathbb{E}_{E^i_{j|\mathbf{h}}\sim U(\pm\frac{q}{2})}\left[e^{-SE^i_{j|\mathbf{h}}}\right] \tag{76}$$

$$= \int_{-q/2}^{q/2} e^{Sx}\frac{1}{q}\mathrm{d}x = \frac{e^{Sq/2}-e^{-Sq/2}}{Sq}. \tag{77}$$

Thus, similar as in the proof of Theorem 2, we obtain

$$\mathrm{P}\left(\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T(\mathbf{W}/S + \mathbf{E}) \leq 0\right) \tag{78}$$

$$\leq \prod_{(i,j,\mathbf{h})\in A_1} \mathbb{E}\left[e^{-W^i_{j|\mathbf{h}}}\right]\mathbb{E}\left[e^{-SE^i_{j|\mathbf{h}}}\right] \cdot \prod_{(i,j,\mathbf{h})\in A_2} \mathbb{E}\left[e^{W^i_{j|\mathbf{h}}}\right]\mathbb{E}\left[e^{SE^i_{j|\mathbf{h}}}\right]. \tag{79}$$

Consequently,

$$\sum_{c' \neq c} \mathrm{P}\left(\left(\boldsymbol{\phi}(c,\mathbf{x}) - \boldsymbol{\phi}(c',\mathbf{x})\right)^T (\mathbf{W}/S + \mathbf{E}) \leq 0\right) \tag{80}$$
$$\leq \sum_{c' \neq c^{(n)}} F^{2(L+1)} \left[\prod_{(i,j,\mathbf{h}) \in A_1} M^i_{j|\mathbf{h}}\right] \left[\prod_{(i,j,\mathbf{h}) \in A_2} \widetilde{M}^i_{j|\mathbf{h}}\right],$$

where $F = \frac{e^{Sq/2} - e^{-Sq/2}}{Sq}$. ■

IV. Experiments

The derived bounds are evaluated on real-world datasets. Additionally, we compare generatively and discriminatively optimized BNCs with respect to *classification performance* and *robustness* against parameter quantization. The investigation of classification performance compares the absolute CRs of generatively and discriminatively optimized BNCs with respect to parameter quantization, while the investigation of robustness compares the number of changing classifications due to parameter quantization of these two types of BNCs.

In all experiments, we select the scale factor S as the minimum value such that all parameters can be represented using B bits, cf. Section III-B.

A. Data Sets

We perform experiments for USPS, MNIST and DC-Mall data. In the following, we provide details about the data:

a) USPS Data [27]: This dataset contains 11000 uniformly distributed handwritten digit images from zip codes of mail envelopes. Each digit is represented as a 16×16 grayscale image, where each pixel is considered as feature. 8000 samples are used for training and 3000 for testing.

b) MNIST Data [28]: This dataset contains 70000 samples of handwritten digits, i.e. 7000 samples of each digit. 60000 samples are used for training and 10000 for testing. The digits represented by gray-level images were down-sampled by a factor of two resulting in a resolution of 16×16 pixels, i.e. 196 features.

c) DC-Mall Data [4]: This dataset contains a hyper-spectral remote sensing image of the Washington D.C. Mall area. In total, there are 1280×307 hyper-spectral pixels, each containing 191 spectral bands. From these spectral bands, individual pixels are to be classified to one of 7 classes (roof, road, grass, trees, trail, water, or shadow). For each class, 5000 samples are used for training, i.e. in total 35000 samples. The remaining 357960 samples are used for testing.

B. Classification Performance Bounds

We evaluate the bounds derived in Section III. For all datasets, we perform Bayesian parameter estimation, where we assume a uniform parameter prior, i.e. $\widetilde{\alpha}^i_{j|\mathbf{h}} = 1$. For $B \in \{1, \ldots, 10\}$ bits we determine the scale factor as stated above and quantize the MAP parameters. BNCs using the resulting reduced precision parameters are evaluated by computing the CR performance on the test set. Additionally, the bounds on the CR derived in the previous section are computed. As reference, we also determine the CR performance of BNCs using double-precision MAP parameters. Furthermore, we compare BNCs with NB and generatively optimized TAN structures. These TAN structures are learned using the conditional mutual information (TAN-CMI) criterion [3]. We did not consider more general structures, e.g. 2-trees for augmenting NB, because 1) there is no significant increase in classification performance on several standard benchmarks for these structures [11], 2) inference complexity scales with the tree-width of the corresponding moralized graph, i.e. computational complexity increases drastically which conflicts with our interest for computationally highly efficient models, and 3) more general models have more parameters to be estimated (from the same number of samples) and therefore parameter estimates have a higher variance.

TABLE I
Number of parameters of BNCs for different datasets and structures

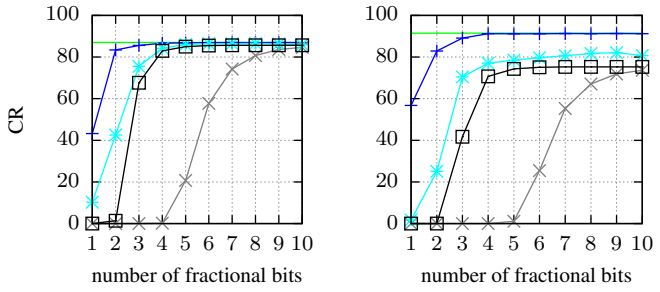
Dataset	Structure	#Parameters	Samples/Parameter
USPS	NB	8650	0.92
	TAN-CMI	33040	0.24
	TAN-MM	31320	0.26
MNIST	NB	6720	8.93
	TAN-CMI	38350	1.56
	TAN-MM	39370	1.52
DC-Mall	NB	21490	12.19
	TAN-CMI	574406	0.46
	TAN-MM	484813	0.54

Experimental results are shown in Figure 2. CR performance of the reduced precision BNCs increases with the number of bits and is close to optimal when using 3 to 4 bits for all datasets[5]. The worst-case bounds are conservative and rather loose. In contrast, the probabilistic bounds are much tighter. The bounds derived assuming uniform and independent quantization errors, cf. Theorem 3, are less tight than those obtained by assuming beta distributed parameters, cf. Theorem 2.

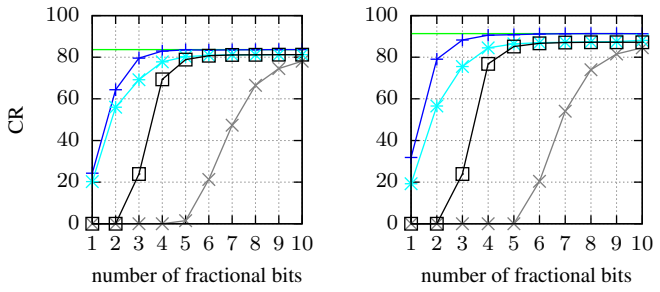
While BNCs with TAN structures typically achieve better CRs, the determined bounds for TAN structures are looser than the bounds for NB structures. This is because in the case of TAN structures, less samples are available for estimating the CPTs, cf. Table I. Therefore, parameter estimates are more uncertain and quantization effects can have larger impact. This effect is strongest for the USPS dataset which consists only of a small number of training samples, i.e. the number of samples per parameter, as shown in Table I, is low.

To emphasize the dependence of the bounds on the sample size, we performed the following experiment: For MNIST we trained a BNC with TAN-CMI structure on $10\%, 20\%, 50\%$ and 100% of the samples and computed the performance bounds, respectively. The results are shown in Figure 3. With increasing sample size, the bounds become tighter because the variance of the parameter estimates reduces.

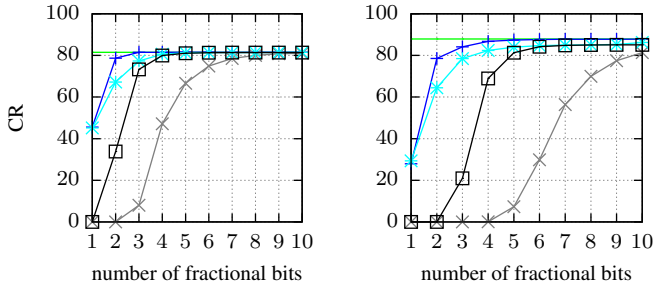
[5]Classification requires adding up $L+1$ reduced precision parameters and performing a maximum operation. Hence, additional $\log_2(L+1)$ bits are required to avoid an overflow during summation. For easier comparison across different datasets, these additional bits are not considered in the presented figures.



(a) MAP parameters, USPS, left: NB, right: TAN-CMI



(b) MAP parameters, MNIST, left: NB, right: TAN-CMI



(c) MAP parameters, DC-Mall, left: NB, right: TAN-CMI

Fig. 2. CRs and bounds on the CRs of BNCs with reduced precision parameters for varying number of bits; worst-case bounds (grey $\times$), probabilistic bound (cyan $\star$), probabilistic bound assuming uniform quantization error (black $\square$), reduced precision MAP parameters (blue $+$), full-precision MAP parameters (green, no marker).

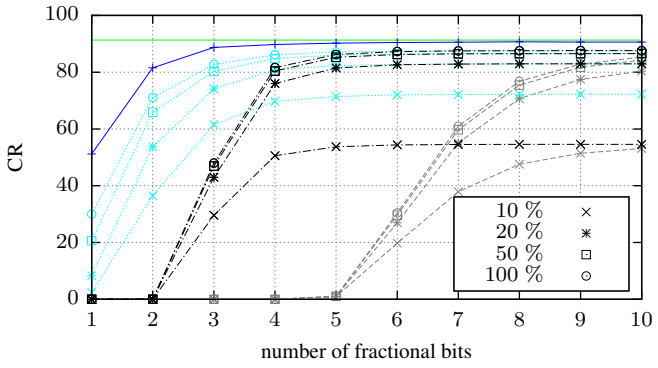


Fig. 3. Sample size dependence of the proposed performance bounds for MNIST data and BNCs with TAN-CMI structure; worst-case bounds (dashed grey), probabilistic bounds (dotted cyan), probabilistic bounds assuming uniform quantization error (dash-dotted black), reduced precision MAP parameters (solid blue), full-precision MAP parameters (green, horizontal); bounds are learned on 10%, 20%, 50% and 100% of the training data (from bottom to top).

TABLE II
COMPARISON OF THE CRS OF BNCS WITH MAP AND MM PARAMETERS; A PLUS (MINUS) SIGN INDICATES THAT FOR THE CORRESPONDING DATASET/STRUCTURE/NUMBER OF BITS BNCS WITH MM (MAP) PARAMETERS HAVE A SIGNIFICANTLY HIGHER CR

		number of bits									
Dataset	Structure	1	2	3	4	5	6	7	8	9	10
USPS	NB	−	−	+	+	+	+	+	+	+	+
	TAN-CMI	−	−	+	+	+	+	+	+	+	+
MNIST	NB	−	−	+	+	+	+	+	+	+	+
	TAN-CMI	−	−	+	+	+	+	+	+	+	+
DC-Mall	NB	−	+	−	+	+	+	+	+	+	+
	TAN-CMI	−	−	+	+	+	+	+	+	+	+

C. Classification Performance of BNCs Optimized for Large Margin

We compare the classification performance of BNCs with generatively and discriminatively optimized parameters/structure with respect to parameter quantization. One motivation for this is that determining whether parameter quantization changes the decision of a BNC $\mathcal{B}$ for a sample belonging to class c with feature instantiation $\mathbf{x}$ or not, is equivalent to determining whether the error introduced by quantization is larger than its log margin, i.e.

$$\log d = \log \mathrm{P}^{\mathcal{B}}(c, \mathbf{x}) - \max_{c' \neq c} \log \mathrm{P}^{\mathcal{B}}(c', \mathbf{x}). \tag{81}$$

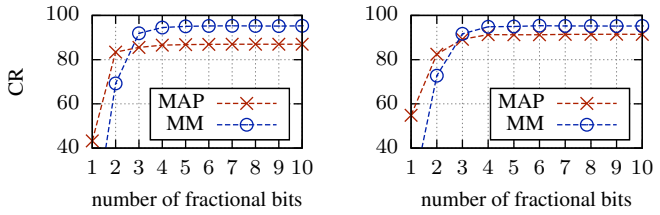
Maximizing this margin is essentially the objective of large margin training of BNCs, cf. [4], [23], [12], [29]. Hence, the assumption that BNCs optimized for a large margin are more *robust* to parameter quantization than for other BNC learning approaches is obtruding (this assumption is also supported by experimental results presented in [1], [2]). In the following, we compare two different approaches for obtaining large margin BNCs.

a) Discriminatively versus generatively optimized parameters: Here, we compare the classification performance of BNCs with MAP parameters and of BNCs with MM parameters over varying numbers of bits used for quantization. MM parameters are determined using the algorithm described in [4]. The structures considered are NB and TAN-CMI. The results for USPS, MNIST and DC-Mall data are shown in Figure 4. Our hypothesis is that the classification rate of BNCs with MM parameters is higher than that of BNCs with MAP parameters, i.e.

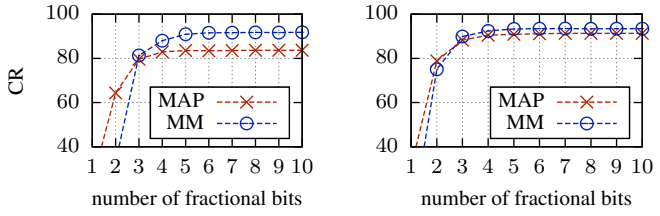
$$\mathrm{CR}\big(h_{\mathcal{Q}(\mathbf{w}^{\mathrm{MM}})}\big) \geq \mathrm{CR}\big(h_{\mathcal{Q}(\mathbf{w}^{\mathrm{MAP}})}\big), \tag{82}$$

where, in abuse of notation, $h_{\mathcal{Q}(\mathbf{w})}$ is the BNC induced by the quantized parameters $\mathcal{Q}(\mathbf{w})$. We use a one-tailed dependent t-test for paired samples at significance level 0.01 for testing significance. Results are summarized in Table II. For all but small bit-widths and all datasets and structures, discriminatively optimized parameters yield significantly higher CRs.

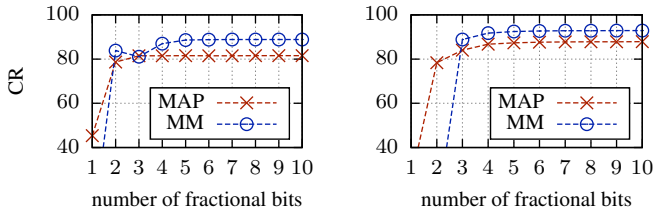
b) Discriminatively optimized BN structures: We compare the CR performance of BNCs with MAP parameters using TAN-CMI and margin-optimized TAN structures (TAN-MM). TAN-MM structures are determined using the margin



(a) USPS, NB, left: NB, right: TAN-CMI



(b) MNIST, left: NB, right: TAN-CMI



(c) DC-Mall, left: NB, right: TAN-CMI

Fig. 4. CRs of BNCs with MAP and MM parameters over varying number of bits; MAP parameters (red); MM parameters (blue)

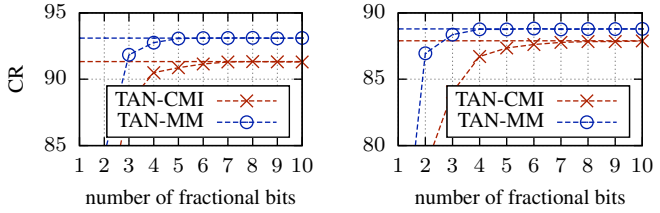


Fig. 5. CRs of BNCs with MAP parameters over varying number of bits, TAN-CMI and TAN-MM structures; Left: MNIST, Right: DC-Mall.

objective (11) embedded in a hinge loss function as objective for scoring BN structures. Optimization is performed using a greedy hill climbing heuristic. Details are provided in [29]. CR results for MNIST and DC-Mall data are shown in Figure 5. Formally, our hypothesis is that the classification rate of BNCs with TAN-MM structure is higher than that of BNCs with TAN-CMI structures, i.e.

$$\mathrm{CR}\big(h_{\mathcal{Q}(\mathbf{w}^{\text{TAN-MM}})}\big) \geq \mathrm{CR}\big(h_{\mathcal{Q}(\mathbf{w}^{\text{TAN-CMI}})}\big), \tag{83}$$

where $\mathbf{w}^{\text{TAN-MM}}$ denotes the MAP parameters for the TAN-MM structure and $\mathbf{w}^{\text{TAN-CMI}}$ denotes the MAP parameters for the TAN-CMI structure, respectively. We performed the same statistical tests as above. In all cases, i.e. for all datasets and considered bit-widths (1 to 10 bits), the CR of BNCs with TAN-MM structure is significantly higher.

TABLE III
COMPARISON OF THE ROBUSTNESS OF BNCS WITH MAP AND MM PARAMETERS; A PLUS (MINUS) SIGN INDICATES THAT FOR THE CORRESPONDING DATASET/STRUCTURE/NUMBER OF BITS BNCS WITH MM (MAP) PARAMETERS ARE SIGNIFICANTLY MORE ROBUST

		number of bits									
Dataset	Structure	1	2	3	4	5	6	7	8	9	10
USPS	NB	−	−								
	TAN-CMI	−	−				+				
MNIST	NB	−	−	−	−	−	−				
	TAN-CMI	−	−								
DC-Mall	NB	−	−	−	−	−	−	−	−	−	−
	TAN-CMI	−	−	+	+	+	+	+	+	+	

D. Robustness of BNCs Optimized for Large Margin

We compare the robustness of BNCs with generatively and discriminatively optimized parameters and structure with respect to parameter quantization. We denote a classifier as robust if only a small number of classifications change due to parameter quantization — this is formalized below. The assumption, that BNCs with parameters/structures optimized for a large margin are more robust than other BNCs seems likely (this assumption is also supported by experimental results presented in [1], [2]). However, this higher robustness cannot be observed empirically in all cases. Again, we compare two different approaches for obtaining large margin BNCs.

a) Discriminatively versus generatively optimized parameters: For testing robustness, our hypothesis is that

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}\left(h_{\mathbf{w}^{\text{MM}}}(\mathbf{X}) = h_{\mathcal{Q}(\mathbf{w}^{\text{MM}})}(\mathbf{X})\right)\right] \\ &\quad\geq \mathbb{E}\left[\mathbf{1}\left(h_{\mathbf{w}^{\text{MAP}}}(\mathbf{X}) = h_{\mathcal{Q}(\mathbf{w}^{\text{MAP}})}(\mathbf{X})\right)\right].\end{aligned} \tag{84}$$

Significance of results is assessed using a dependent t-test for paired samples at significance level 0.01. Then, BNCs with discriminatively optimized parameters are almost never significantly more robust to parameter quantization, cf. Table III (the only exception is BNCs with TAN-CMI structure for DC-Mall data). This can be explained with results from sensitivity analysis, cf. Section I — discriminatively optimized BNCs have more extreme parameters, i.e. parameters close to zero or one, than generatively optimized BNCs [1]. Nevertheless, using only 3 to 4 bits, the discriminatively optimized parameters yield higher absolute CRs, cf. Section IV-C and Table II.

b) Discriminatively optimized BN structures: Our hypothesis is that BNCs with large-margin structures are more robust to quantization, i.e.

$$\begin{aligned}&\mathbb{E}\left[\mathbf{1}\left(h_{\mathbf{w}^{\text{TAN-MM}}}(\mathbf{X}) = h_{\mathcal{Q}(\mathbf{w}^{\text{TAN-MM}})}(\mathbf{X})\right)\right] \\ &\quad\geq \mathbb{E}\left[\mathbf{1}\left(h_{\mathbf{w}^{\text{TAN-CMI}}}(\mathbf{X}) = h_{\mathcal{Q}(\mathbf{w}^{\text{TAN-CMI}})}(\mathbf{X})\right)\right].\end{aligned} \tag{85}$$

For assessing results, we used the same statistical test as above. BNCs with margin-optimized structures show a higher robustness to parameter quantization than TAN-CMI structures, cf. Table IV. Additionally, the CR performance using margin-optimized structures is always better, cf. Section IV-C and Figure 5. Hence, in this case the large margin structures seem favorable.

TABLE IV
COMPARISON OF THE ROBUSTNESS OF BNCS WITH TAN-CMI AND TAN-MM STRUCTURES AND MAP PARAMETERS; A PLUS SIGN INDICATES THAT FOR THE CORRESPONDING DATASET/NUMBER OF BITS BNCS WITH TAN-MM STRUCTURE ARE SIGNIFICANTLY MORE ROBUST

	number of bits									
Dataset	1	2	3	4	5	6	7	8	9	10
USPS	+	+	+	+	+					
MNIST	+	+	+	+	+					
DC-Mall	+	+	+	+	+	+	+	+	+	+

V. CONCLUSION AND FUTURE WORK

We considered BNCs with reduced precision parameters and derived an easy to evaluate worst-case bound on the CR performance. Furthermore, a probabilistic bound on the CR and approximations for the expected value of the quantized parameters were derived. In experiments, we evaluated the performance of reduced precision BNCs and the derived bounds. Only 3 to 4 bits for representing each parameter are necessary to achieve CR performance close to double-precision floating-point performance. We investigated classification performance and robustness of BNCs with generatively and discriminatively optimized parameters and structures with respect to parameter quantization. While discriminatively optimized parameters do not show higher robustness than generatively optimized parameters, the CR performance of BNCs in the former case is already better when using only 2 to 3 bits for representing each parameter. When using discriminatively optimized TAN-MM structures, robustness to parameter quantization as well as CR performance is increased.

Future work includes several directions: 1) While in this work parameters were initially learned in full-precision and subsequently quantized, we aim at learning parameters using reduced-precision computations only. In this way, reduced-precision BNCs could be used in online scenarios. However, this raises the need for specialized algorithms and careful investigation of effects caused by reduced-precision computations. 2) We work towards a real-world reduced-precision implementation of BNCs that could be used in hearing aids for acoustic scene classification. Once the implementation is completed, we want to compare it to full-precision implementations with respect to speedup, power consumption and classification performance. 3) We are interested in learning optimal parameters given a specified number of bits, where optimality is measured by a generative or discriminative criterion (partial results in this regard are already available [10]). In both cases, we want to derive bounds relating the performance obtained when using optimal parameters to the performance obtained when using rounded parameters.

REFERENCES

[1] S. Tschiatschek, P. Reinprecht, M. Mücke, and F. Pernkopf, "Bayesian network classifiers with reduced precision parameters," in *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2012, pp. 74–89.

[2] S. Tschiatschek, C. E. Cancino Chacón, and F. Pernkopf, "Bounds for Bayesian network classifiers with reduced precision parameters," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 3357–3361.

[3] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, pp. 131–163, 1997.

[4] F. Pernkopf, M. Wohlmayr, and S. Tschiatschek, "Maximum margin Bayesian network classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 3, pp. 521–531, 2012.

[5] D. Husmeier, R. Dybowski, and S. Roberts, *Probabilistic Modelling in Bioinformatics and Medical Informatics*. Springer, 2004.

[6] P. Helman, R. Veroff, S. R. Atlas, and C. Willman, "A Bayesian network classification methodology for gene expression data," *Computational Biology*, vol. 11, 2004.

[7] D.-U. Lee, A. Gaffar, R. C. C. Cheung, O. Mencer, W. Luk, and G. Constantinides, "Accuracy-guaranteed bit-width optimization," *IEEE Transacations on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, pp. 1990–2000, 2006.

[8] H. Chan and A. Darwiche, "When do numbers really matter?" *Artcial Intelligence Research*, vol. 17, no. 1, pp. 265–287, 2002.

[9] F. Pernkopf, M. Wohlmayr, and M. Mücke, "Maximum margin structure learning of Bayesian network classifiers," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 2076–2079.

[10] S. Tschiatschek, K. Paul, and F. Pernkopf, "Integer Bayesian network classifiers," in *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, 2014.

[11] F. Pernkopf and J. Bilmes, "Efficient heuristics for discriminative structure learning of bayesian network classifiers," *Journal of Machine Learning Research*, vol. 11, pp. 2323–2360, Aug. 2010.

[12] R. Peharz and F. Pernkopf, "Exact maximum margin structure learning of Bayesian networks," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[13] R. Greiner and W. Zhou, "Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers," in *National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, 2002, pp. 167–173.

[14] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri, "On discriminative Bayesian network classifiers and logistic regression," *Machine Learning*, vol. 59, no. 3, pp. 267–296, 2005.

[15] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. New York: John Wiley & Sons, 1994.

[16] D. Heckerman, D. Geiger, and D. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," Tech. Rep., 1995.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[18] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[19] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[20] I. Kononenko, "Semi-naive Bayesian classifier," in *Proceedings of the European Working Session on Learning on Machine Learning*, ser. EWSL-91, 1991, pp. 206–219.

[21] S. Acid, L. M. Campos, and J. G. Castellano, "Learning Bayesian network classifiers: Searching in a space of partially directed acyclic graphs," *Machine Learning*, vol. 59, pp. 213–235, 2005.

[22] F. Pernkopf, R. Peharz, and S. Tschiatschek, "Introduction to probabilistic graphical models," in *Academic Press Library in Signal Processing*, 2014, vol. 1, ch. 18, pp. 989–1064.

[23] Y. Guo, D. Wilkinson, and D. Schuurmans, "Maximum margin Bayesian networks," in *Uncertainty in Artificial Intelligence (UAI)*, 2005, pp. 233–242.

[24] N. Piatkowski, L. Sangkyun, and K. Morik, "The integer approximation of undirected graphical models," in *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, 2014.

[25] B. Widrow, I. Kollar, and M. Liu, "Statistical theory of quantization," *IEEE Transactions on Instrumentation and Measurement*, vol. 45, no. 2, pp. 353–361, 1996.

[26] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. Mc-Graw Hill, 1984.

[27] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, Aug. 2003.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[29] F. Pernkopf and M. Wohlmayr, "Stochastic margin-based structure learning of Bayesian network classifiers," *Pattern Recognition*, vol. 46, no. 2, pp. 464–471, 2013.

Sebastian Tschiatschek received the BSc degree and MSc degree (with distinction) in Electrical Engineering at Graz University of Technology (TUG) in 2007 and 2010, respectively. He conducted his Master thesis during a one-year stay at ETH Zürich, Switzerland. Currently, he is with the Signal Processing and Speech Communication Laboratory at TUG where he is pursuing the PhD degree. His research interests include Bayesian networks, information theory in conjunction with graphical models and statistical pattern recognition.

Franz Pernkopf received his MSc (Dipl. Ing.) degree in Electrical Engineering at Graz University of Technology, Austria, in summer 1999. He earned a PhD degree from the University of Leoben, Austria, in 2002. In 2002 he was awarded the Erwin Schrödinger Fellowship. He was a Research Associate in the Department of Electrical Engineering at the University of Washington, Seattle, from 2004 to 2006. Currently, he is Associate Professor at the Laboratory of Signal Processing and Speech Communication, Graz University of Technology, Austria. His research interests include machine learning, discriminative learning, graphical models, feature selection, finite mixture models, and image- and speech processing applications.