Inner Product Spaces for Bayesian Networks

Atsuyoshi Nakamura
Graduate School of Engineering
Hokkaido University, Sapporo 060-8628, Japan
atsu@main.eng.hokudai.ac.jp

Michael Schmitt*, Niels Schmitt, and Hans Ulrich Simon
Fakultät für Mathematik
Ruhr-Universität Bochum, 44780 Bochum, Germany
{mschmitt,nschmitt,simon}@lmi.ruhr-uni-bochum.de

Abstract

Bayesian networks have become one of the major models for statistical inference. We study the question whether the decisions computed by a Bayesian network can be represented within a low-dimensional inner product space. We focus on two-label classification tasks over the Boolean domain. As the main results, we establish upper and lower bounds on the dimension of the "natural" inner product space for Bayesian networks with an explicitly given (full or reduced) parameter collection. In particular, these bounds are tight up to a factor of 2. For some nontrivial cases, we even determine the exact values of this dimension. Further, we consider a variant of the logistic autoregressive Bayesian network and show that every sufficiently expressive inner product space must have dimension at least $2^{\Omega(n)}$, where n is the number of network nodes. As a major technical contribution, this work reveals combinatorial and algebraic structures within Bayesian networks such that known methods for the derivation of lower bounds on the dimension of inner product spaces can be brought into play.

Keywords: Bayesian network, inner product space, embedding, linear arrangement, Euclidean dimension

*Corresponding author

1 Introduction

During the last decade, there has been remarkable interest in learning systems based on hypotheses that can be written as inner products in an appropriate feature space and learned by algorithms that perform a kind of empirical or structural risk minimization. Often in such systems, the inner product operation is not carried out explicitly, but reduced to the evaluation of a so-called kernel function that operates on instances of the original data space. A major advantage of this technique is that it allows to handle high-dimensional feature spaces efficiently. The learning strategy proposed by Boser et al. (1992) in connection with the so-called Support Vector Machine is a theoretically well founded and very powerful method that, in the years since its introduction, has already outperformed most other systems in a wide variety of applications (see also Vapnik, 1998).

Bayesian networks have a long history in statistics. In the first half of the 1980s they were introduced to the field of expert systems through work by Pearl (1982) and Spiegelhalter and Knill-Jones (1984). Bayesian networks are much different from kernel-based learning systems and offer some complementary advantages. They graphically model conditional independence relationships between random variables. Like other probabilistic models, Bayesian networks can be used to represent inhomogeneous data with possibly overlapping features and missing values in a uniform manner. Quite elaborate methods dealing with Bayesian networks have been developed for solving problems in pattern classification.

Recently, several research groups considered the possibility of combining the key advantages of probabilistic models and kernel-based learning systems. For this purpose, various kernels were extensively studied, for instance, by Jaakkola and Haussler (1999a,b), Oliver et al. (2000), Saunders et al. (2003), Tsuda and Kawanabe (2002), and Tsuda et al. (2002, 2004). Altun et al. (2003) proposed a kernel for the Hidden Markov Model, which is a special case of a Bayesian network. A unifying approach for kernel methods and Markov networks has been presented by Taskar et al. (2004).

In this article, we consider Bayesian networks as computational models that perform two-label classification tasks over the Boolean domain. We aim at finding the simplest inner product space that is able to express the concept class, that is, the class of decision functions, induced by a given Bayesian network. Hereby, "simplest" refers to a space which has as few dimensions as possible. We focus on Euclidean spaces equipped with the standard dot product. For the finite-dimensional case, this is no loss of generality since any finite-dimensional reproducing kernel Hilbert space is isometric with $\mathbb{R}^d$ for some d. Furthermore, we use the Euclidean dimension of the space as the measure of complexity. This is well motivated by the fact that most generalization error bounds for linear classifiers are given in terms of either the Euclidean dimension or in terms of

the geometrical margin between the data points and the separating hyperplanes. Applying random projection techniques from Johnson and Lindenstrauss (1984), Frankl and Maehara (1988), or Arriaga and Vempala (1999), it can be shown that any arrangement with a large margin can be converted into a low-dimensional arrangement. A recent result of Balcan et al. (2004) in this direction even takes into account low-dimensional arrangements that allow some error. Thus, a large lower bound on the smallest possible dimension rules out the possibility that a classifier with a large margin exists. Given a Bayesian network $\mathcal{N}$, we introduce $\mathrm{Edim}(\mathcal{N})$ for denoting the smallest dimension d such that the decisions represented by $\mathcal{N}$ can be implemented as inner products in the d-dimensional Euclidean space. Our results are given as upper and lower bounds for $\mathrm{Edim}(\mathcal{N})$.

We first consider Bayesian networks with an explicitly given parameter collection. The parameters can be arbitrary, where we speak of an unconstrained network, or they may be required to satisfy certain restrictions, in which case we have a network with a reduced parameter collection. For both network types, we show that the "natural" inner product space, which can obtained from the probabilistic model by straightforward algebraic manipulations, has a dimension that is the smallest possible up to a factor of 2, and even up to an additive term of 1 in some cases. Furthermore, we determine the exact values of $\mathrm{Edim}(\mathcal{N})$ for some nontrivial instances of these networks. The lower bounds in all these cases are are obtained by analyzing the Vapnik-Chervonenkis (VC) dimension of the concept class associated with the Bayesian network. Interestingly, the VC dimension plays also a major role when estimating the sample complexity of a learning system. In particular, it can be used to derive bounds on the number of training examples that are required for selecting hypotheses that generalize well on new data. Thus, the tight bounds on $\mathrm{Edim}(\mathcal{N})$ reveal that the smallest possible Euclidean dimension for a Bayesian network with an explicitly given parameter collection is closely tied to its sample complexity.

Second, we investigate a class of probabilistic models known as logistic autoregressive Bayesian networks, also called sigmoid belief networks. These networks were originally proposed by McCullagh and Nelder (1983) and studied systematically, for instance, by Neal (1992), and Saul et al. (1996). (See also Frey, 1998). Using the VC dimension, we show that $\mathrm{Edim}(\mathcal{N})$ for theses networks must grow at least as $\Omega(n^2)$, where n is the number of nodes.

Finally, we get interested in the question whether it is possible to establish an exponential lower bound on $\mathrm{Edim}(\mathcal{N})$ for the logistic autoregressive Bayesian network. This investigation is motivated by the fact also derived here that these networks have a VC dimension bounded by $O(n^6)$. Hence, VC dimension considerations were insufficient to yield an exponential lower bound for $\mathrm{Edim}(\mathcal{N})$. We succeed in giving a positive answer for a variant of these networks that we call the modified logistic autoregressive Bayesian network. This variant is also shown to have VC dimension $O(n^6)$. We obtain that for a network with $n+2$ nodes, $\mathrm{Edim}(\mathcal{N})$ is at least as large as $2^{n/4}$. The proof for this lower bound uses

the construction of embedding one concept class into another. In particular, we show that a certain class of Boolean parity functions can be embedded into such a network.

While, as mentioned above, the connection between probabilistic models and inner product spaces has already been investigated, this work seems to be the first one that explicitly addresses the question of finding a smallest-dimensional sufficiently expressive inner product space. It should be mentioned, however, that there has also been research (not concerned with probabilistic models) considering the related question of representing a given concept class by a system of halfspaces (Ben-David et al., 2002; Forster et al., 2001; Forster, 2002; Forster and Simon, 2002; Forster et al., 2003; Kiltz, 2003; Kiltz and Simon, 2003). A further contribution of our work can be seen in the uncovering of combinatorial and algebraic structures within Bayesian networks such that techniques known from this literature can be brought into play.

We start by introducing the basic concepts in Section 2. The upper bounds are presented in Section 3. Section 4 deals with lower bounds that are obtained using the VC dimension as the central tool. The exponential lower bound for the modified logistic autoregressive network is derived in Section 5. In Section 6 we draw the major conclusions and mention some open problems.

Bibliographic Note. Results in this article have been presented at the 17th Annual Conference on Learning Theory, COLT 2004, in Banff, Canada (Nakamura et al., 2004).

2 Preliminaries

In the following, we give formal definitions for the basic notions in this article. Section 2.1 introduces terminology from learning theory. In Section 2.2, we define Bayesian networks and the distributions and concept classes they induce. The idea of a linear arrangement for a concept class is presented in Section 2.3.

2.1 Concept Classes, VC Dimension, and Embeddings

A *concept class* $\mathcal{C}$ over domain $\mathcal{X}$ is a family of functions of the form $f : \mathcal{X} \to \{-1, 1\}$. Each $f \in \mathcal{C}$ is called a *concept*. A finite set $S = \{s_1, \ldots, s_m\} \subseteq \mathcal{X}$ is said to be *shattered* by $\mathcal{C}$ if for every binary vector $b \in \{-1, 1\}^m$ there exists some concept $f \in \mathcal{C}$ such that $f(s_i) = b_i$ for $i = 1, \ldots, m$. The *Vapnik-Chervonenkis (VC) dimension* of $\mathcal{C}$ is given by

$$\mathrm{VCdim}(\mathcal{C}) \;=\; \sup\{m \mid \text{there is some } S \subseteq \mathcal{X} \text{ shattered by } \mathcal{C} \text{ and } |S| = m\}.$$

For every $z \in \mathbb{R}$, let $\mathrm{sign}(z) = 1$ if $z \geq 0$, and $\mathrm{sign}(z) = -1$ otherwise. We use the sign function for mapping a real-valued function g to a ± 1-valued concept

sign $\circ\, g$.

Given a concept class $\mathcal{C}$ over domain $\mathcal{X}$ and a concept class $\mathcal{C}'$ over domain $\mathcal{X}'$, we write $\mathcal{C} \leq \mathcal{C}'$ if there exist mappings

$$\mathcal{C} \ni f \mapsto f' \in \mathcal{C}' \quad \text{and} \quad \mathcal{X} \ni x \mapsto x' \in \mathcal{X}'$$

satisfying

$$f(x) = f'(x') \text{ for every } f \in \mathcal{C} \text{ and } x \in \mathcal{X}.$$

These mappings are said to provide an *embedding* of $\mathcal{C}$ into $\mathcal{C}'$. Obviously, if $S \subseteq \mathcal{X}$ is an m-element set that is shattered by $\mathcal{C}$ then $S' = \{s' \mid s \in S\} \subseteq \mathcal{X}'$ is an m-element set that is shattered by $\mathcal{C}'$. Consequently, $\mathcal{C} \leq \mathcal{C}'$ implies $\mathrm{VCdim}(\mathcal{C}) \leq \mathrm{VCdim}(\mathcal{C}')$.

2.2 Bayesian Networks

Definition 1. *A* Bayesian network $\mathcal{N}$ *has the following components:*

1. *A directed acyclic graph $G = (V, E)$, where V is a finite set of nodes and $E \subseteq V \times V$ a set of edges,*

2. *a collection $(p_{i,\alpha})_{i \in V, \alpha \in \{0,1\}^{m_i}}$ of programmable parameters with values in the open interval $]0, 1[$, where m_i denotes the number of predecessors of node i, that is, $m_i = |\{j \in V \mid (j, i) \in E\}|$,*

3. *constraints that describe which assignments of values from $]0, 1[$ to the parameters of the collection are allowed.*

If the constraints are empty, we speak of an unconstrained *network. Otherwise, the network is* constrained.

We identify the $n = |V|$ nodes of $\mathcal{N}$ with the numbers $1, \ldots, n$ and assume that every edge $(j, i) \in E$ satisfies $j < i$, that is, E induces a topological ordering on $\{1, \ldots, n\}$. Given $(j, i) \in E$, j is called a parent of i. We use P_i to denote the set of parents of node i, and let $m_i = |P_i|$ be the number of parents. A network $\mathcal{N}$ is said to be *fully connected* if $P_i = \{1, \ldots, i-1\}$ holds for every node i.

Example 1 (kth-order Markov chain). *For $k \geq 0$, let $\mathcal{N}_k$ denote the unconstrained Bayesian network with $P_i = \{i-1, \ldots, i-k\}$ for $i = 1, \ldots, n$ (with the convention that numbers smaller than 1 are ignored such that $m_i = |P_i| = \min\{i-1, k\}$). The total number of parameters is equal to $2^k(n-k) + 2^{k-1} + \cdots + 2 + 1 = 2^k(n-k+1) - 1$.*

We associate with every node i a Boolean variable x_i with values in $\{0,1\}$. We say x_j is a parent-variable of x_i if j is a parent of i. Each $\alpha \in \{0,1\}^{m_i}$ is called a possible bit-pattern for the parent-variables of x_i. We use $M_{i,\alpha}$ to denote the polynomial

$$M_{i,\alpha}(x) \;=\; \prod_{j \in P_i} x_j^{\alpha_j}, \text{ where } x_j^0 = 1 - x_j \text{ and } x_j^1 = x_j,$$

that is, $M_{i,\alpha}(x)$ is 1 if the parent variables of x_i exhibit bit-pattern α, otherwise it is 0.

Bayesian networks are graphical models of conditional independence relationships. This general idea is made concrete by the following notion.

Definition 2. *Let $\mathcal{N}$ be a Bayesian network with nodes $1, \ldots, n$. The class of distributions induced by $\mathcal{N}$, denoted as $\mathcal{D}_{\mathcal{N}}$, consists of all distributions on $\{0,1\}^n$ of the form*

$$P(x) \;=\; \prod_{i=1}^{n} \prod_{\alpha \in \{0,1\}^{m_i}} p_{i,\alpha}^{x_i M_{i,\alpha}(x)} (1 - p_{i,\alpha})^{(1-x_i) M_{i,\alpha}(x)}. \qquad (1)$$

Thus, for every assignment of values from $]0,1[$ to the parameters of $\mathcal{N}$, we obtain a specific distribution from $\mathcal{D}_{\mathcal{N}}$. Recall that not every possible assignment is allowed if $\mathcal{N}$ is constrained.

The polynomial representation of $\log(P(x))$ resulting from equation (1) is known as "Chow expansion" in the pattern classification literature (see, e.g., Duda and Hart, 1973). The parameter $p_{i,\alpha}$ represents the conditional probability for the event $x_i = 1$ given that the parent variables of x_i exhibit bit-pattern α. Equation (1) is a chain expansion for $P(x)$ that expresses $P(x)$ as a product of conditional probabilities.

An unconstrained network that is highly connected may have a number of parameters that grows exponentially in the number of nodes. The idea of a constrained network is to keep the number of parameters reasonably small even in case of a dense topology. We consider two types of constraints giving rise to the definitions of networks with a reduced parameter collection and logistic autoregressive networks.

Definition 3. *A* Bayesian network with a reduced parameter collection *is a Bayesian network with the following constraints: For every $i \in \{1, \ldots, n\}$ there exists a surjective function $R_i : \{0,1\}^{m_i} \to \{1, \ldots, d_i\}$ such that the parameters of $\mathcal{N}$ satisfy*

$$\forall i = 1, \ldots, n, \forall \alpha, \alpha' \in \{0,1\}^{m_i} : R_i(\alpha) = R_i(\alpha') \implies p_{i,\alpha} = p_{i,\alpha'}.$$

We denote the network as $\mathcal{N}^R$ for $R = (R_1, \ldots, R_n)$. Obviously, $\mathcal{N}^R$ is completely described by the reduced parameter collection $(p_{i,c})_{1 \le i \le n, 1 \le c \le d_i}$.

A special case of these networks uses decision trees or graphs to represent the parameters.

Example 2. *Chickering et al. (1997) proposed Bayesian networks "with local structure". These networks contain a decision tree T_i (or, alternatively, a decision graph G_i) over the parent-variables of x_i for every node i. The conditional probability for $x_i = 1$, given the bit-pattern of the variables from P_i, is attached to the corresponding leaf in T_i (or sink in G_i, respectively). This fits nicely into our framework of networks with a reduced parameter collection. Here, d_i denotes the number of leaves in T_i (or sinks of G_i, respectively), and $R_i(\alpha)$ is equal to $c \in \{1, \ldots, d_i\}$ if α is routed to leaf c in T_i (or to sink c in G_i, respectively).*

For a Bayesian network with reduced parameter collection, the distribution $P(x)$ from Definition 2 can be written in a simpler way. Let $R_{i,c}(x)$ denote the $\{0,1\}$-valued function that indicates for every $x \in \{0,1\}^n$ whether the projection of x to the parent-variables of x_i is mapped by R_i to the value c. Then, we have

$$P(x) \;=\; \prod_{i=1}^{n} \prod_{c=1}^{d_i} p_{i,c}^{x_i R_{i,c}(x)} (1 - p_{i,c})^{(1-x_i) R_{i,c}(x)}. \tag{2}$$

We finally introduce the so-called logistic autoregressive Bayesian networks, originally proposed by McCullagh and Nelder (1983), that have been shown to perform surprisingly well on certain problems (see also Neal, 1992, Saul et al., 1996, and Frey, 1998).

Definition 4. *The* logistic autoregressive Bayesian network $\mathcal{N}_\sigma$ *is the fully connected Bayesian network with constraints on the parameter collection given as*

$$\forall i = 1, \ldots, n, \exists (w_{i,j})_{1 \le j \le i-1} \in \mathbb{R}^{i-1}, \forall \alpha \in \{0,1\}^{i-1} : p_{i,\alpha} = \sigma \left(\sum_{j=1}^{i-1} w_{i,j} \alpha_j \right),$$

where $\sigma(y) = 1/(1 + e^{-y})$ is the standard sigmoid function. Obviously, $\mathcal{N}_\sigma$ is completely described by the parameter collection $(w_{i,j})_{1 \le i \le n, 1 \le j \le i-1}$.

In a two-label classification task, functions $P(x), Q(x) \in \mathcal{D}_\mathcal{N}$ are used as discriminant functions, where $P(x)$ and $Q(x)$ represent the distributions of x conditioned to label 1 and -1, respectively. The corresponding decision function assigns label 1 to x if $P(x) \ge Q(x)$, and -1 otherwise. The obvious connection to concept classes in learning theory is made explicit in the following definition.

Definition 5. *Let $\mathcal{N}$ be a Bayesian network with nodes $1, \ldots, n$ and let $\mathcal{D}_\mathcal{N}$ be the corresponding class of distributions. The class of concepts induced by $\mathcal{N}$, denoted as $\mathcal{C}_\mathcal{N}$, consists of all ± 1-valued functions on $\{0,1\}^n$ of the form* $\mathrm{sign}(\log(P(x)/Q(x)))$ *for $P, Q \in \mathcal{D}_\mathcal{N}$.*

Note that the function $\mathrm{sign}(\log(P(x)/Q(x)))$ attains the value 1 if $P(x) \ge Q(x)$, and the value -1 otherwise. We use $\mathrm{VCdim}(\mathcal{N})$ to denote the VC dimension of $\mathcal{C}_\mathcal{N}$.

2.3 Linear Arrangements in Inner Product Spaces

We are interested in embedding concept classes into finite-dimensional Euclidean spaces equipped with the standard dot product $u^\top v = \sum_{i=1}^{d} u_i v_i$, where $u^\top$ denotes the transpose of u. Such an embedding is provided by a linear arrangement. Given a concept class $\mathcal{C}$, we aim at determining the smallest Euclidean dimension, denoted $\mathrm{Edim}(\mathcal{C})$, that such a space can have.

Definition 6. *A* d-dimensional linear arrangement *for a concept class* $\mathcal{C}$ *over domain* $\mathcal{X}$ *is given by collections* $(u_f)_{f\in\mathcal{C}}$ *and* $(v_x)_{x\in\mathcal{X}}$ *of vectors in* $\mathbb{R}^d$ *such that*

$$\forall f \in \mathcal{C}, x \in \mathcal{X} : f(x) = \mathrm{sign}(u_f^\top v_x).$$

The smallest d *such that there exists a* d*-dimensional linear arrangement for* $\mathcal{C}$ *is denoted as* $\mathrm{Edim}(\mathcal{C})$*. If there is no finite-dimensional linear arrangement for* $\mathcal{C}$*,* $\mathrm{Edim}(\mathcal{C})$ *is defined to be infinite.*

If $\mathcal{C}_\mathcal{N}$ is the concept class induced by a Bayesian network $\mathcal{N}$, we write $\mathrm{Edim}(\mathcal{N})$ instead of $\mathrm{Edim}(\mathcal{C}_\mathcal{N})$. It is evident that $\mathrm{Edim}(\mathcal{C}) \leq \mathrm{Edim}(\mathcal{C}')$ if $\mathcal{C} \leq \mathcal{C}'$.

It is easy to see that $\mathrm{Edim}(\mathcal{C}) \leq \min\{|\mathcal{C}|, |\mathcal{X}|\}$ for finite concept classes. Nontrivial upper bounds on $\mathrm{Edim}(\mathcal{C})$ are usually obtained constructively by presenting an appropriate arrangement. As for lower bounds, the following result is immediate from a result by Dudley (1978) which states that $\mathrm{VCdim}(\{\mathrm{sign} \circ f \mid f \in \mathcal{F}\}) = d$ for every d-dimensional vector space $\mathcal{F}$ consisting of real-valued functions (see also Anthony and Bartlett, 1999, Theorem 3.5).

Lemma 1. *Every concept class* $\mathcal{C}$ *satisfies* $\mathrm{Edim}(\mathcal{C}) \geq \mathrm{VCdim}(\mathcal{C})$.

Let PARITY_n be the concept class $\{h_a \mid a \in \{0,1\}^n\}$ of parity functions on the Boolean domain given by $h_a(x) = (-1)^{a^\top x}$. The following lower bound, which will be useful in Section 5, is due to Forster (2002).

Corollary 1. $\mathrm{Edim}(\mathrm{PARITY}_n) \geq 2^{n/2}$.

3 Upper Bounds on the Dimension of Inner Product Spaces for Bayesian Networks

This section is concerned with the derivation of upper bounds on $\mathrm{Edim}(\mathcal{N})$. We obtain bounds for unconstrained networks and for networks with a reduced parameter collection by providing concrete linear arrangements. Given a set M, let 2^M denote its power set.

Theorem 1. *Every unconstrained Bayesian network* $\mathcal{N}$ *satisfies*

$$\mathrm{Edim}(\mathcal{N}) \;\leq\; \left|\bigcup_{i=1}^{n} 2^{P_i \cup \{i\}}\right| \;\leq\; 2 \cdot \sum_{i=1}^{n} 2^{m_i}.$$

Proof. From the expansion of P in equation (1) and the corresponding expansion of Q (with parameters $q_{i,\alpha}$ in the role of $p_{i,\alpha}$), we obtain

$$\log \frac{P(x)}{Q(x)} \;=\; \sum_{i=1}^{n} \sum_{\alpha \in \{0,1\}^{m_i}} \left(x_i M_{i,\alpha}(x) \log \frac{p_{i,\alpha}}{q_{i,\alpha}} + (1 - x_i) M_{i,\alpha}(x) \log \frac{1 - p_{i,\alpha}}{1 - q_{i,\alpha}} \right). \tag{3}$$

On the right-hand side of equation (3), we find the polynomials $M_{i,\alpha}(x)$ and $x_i M_{i,\alpha}(x)$. Note that $|\cup_{i=1}^{n} 2^{P_i \cup \{i\}}|$ equals the number of monomials that occur when we express these polynomials as sums of monomials by successive applications of the distributive law. A linear arrangement of the claimed dimensionality is now obtained in the obvious fashion by introducing one coordinate per monomial. □

This result immediately yields an upper bound for Markov chains of order k.

Corollary 2. *Let $\mathcal{N}_k$ denote the kth-order Markov chain given in Example 1. Then,*

$$\mathrm{Edim}(\mathcal{N}_k) \;\leq\; (n - k + 1)2^k.$$

Proof. Apply Theorem 1 and observe that

$$\bigcup_{i=1}^{n} 2^{P_i \cup \{i\}} \;=\; \bigcup_{i=k+1}^{n} \{J_i \cup \{i\} \mid J_i \subseteq \{i-1, \ldots, i-k\}\} \cup \{J \mid J \subseteq \{1, \ldots, k\}\}.$$

□

Similar techniques as used in the proof of Theorem 1 lead to an upper bound for networks with a reduced parameter collection.

Theorem 2. *Let $\mathcal{N}^R$ be a Bayesian network that has a reduced parameter collection $(p_{i,c})_{1 \leq i \leq n, 1 \leq c \leq d_i}$ in the sense of Definition 3. Then,*

$$\mathrm{Edim}(\mathcal{N}^R) \;\leq\; 2 \cdot \sum_{i=1}^{n} d_i.$$

Proof. Recall that the distributions from $\mathcal{D}_{\mathcal{N}^R}$ can be written as in equation (2). We make use of the obvious relationship

$$\log \frac{P(x)}{Q(x)} \;=\; \sum_{i=1}^{n} \sum_{c=1}^{d_i} \left(x_i R_{i,c}(x) \log \frac{p_{i,c}}{q_{i,c}} + (1 - x_i) R_{i,c}(x) \log \frac{1 - p_{i,c}}{1 - q_{i,c}} \right). \tag{4}$$

A linear arrangement of the appropriate dimension is now obtained by introducing two coordinates per pair (i, c): If x is mapped to v_x in this arrangement, then the projection of v_x to the two coordinates corresponding to (i, c) is $(R_{i,c}(x), x_i R_{i,c}(x))$; the appropriate mapping $(P, Q) \mapsto u_{P,Q}$ in this arrangement is easily derived from (4). □

In Section 4 we shall show that the bounds established by Theorem 1 and Theorem 2 are tight up to a factor of 2 and, in some cases, even up to an additive constant of 1.

The linear arrangements for unconstrained Bayesian networks or for Bayesian networks with a reduced parameter collection were easy to find. This is no accident as this holds for every class of distributions (or densities) from the so-called exponential family because (as pointed out, for instance, in Devroye et al., 1996) the corresponding Bayes rule takes a form known as generalized linear rule. From this representation a linear arrangement is evident. Note, however, that the bound given in Theorem 1 is slightly stronger than the bound obtained from the general approach for members of the exponential family.

4 Lower Bounds Based on VC Dimension Considerations

In this section, we derive lower bounds on $\mathrm{Edim}(\mathcal{N})$ that come close to the upper bounds obtained in the previous section. Before presenting the main results in Section 4.2 as Corollaries 3, 5, and Theorem 7, we focus on some specific Bayesian networks for which we determine the exact values of $\mathrm{Edim}(\mathcal{N})$.

4.1 Optimal Bounds for Specific Networks

In the following we calculate exact values of $\mathrm{Edim}(\mathcal{N})$ by establishing lower bounds of $\mathrm{VCdim}(\mathcal{N})$ and applying Lemma 1. This gives us also the exact value of the VC dimension for the respective networks. We recall that $\mathcal{N}_k$ is the kth-order Markov chain defined in Example 1. The concept class arising from network $\mathcal{N}_0$, which we consider first, is the well-known Naïve Bayes classifier.

Theorem 3.

$$\mathrm{Edim}(\mathcal{N}_0) \;=\; \begin{cases} n+1 & \text{if } n \geq 2, \\ 1 & \text{if } n = 1. \end{cases}$$

The proof of this theorem relies on the following result.

Lemma 2. *For every $p, q \in]0,1[$ there exist $w \in \mathbb{R}$ and $b \in]0,1[$ such that*

$$\forall x \in \mathbb{R}: \; x \log \frac{p}{q} + (1-x) \log \frac{1-p}{1-q} = w(x-b) \tag{5}$$

holds. Conversely, for every $w \in \mathbb{R}$ and $b \in]0,1[$ there exist $p, q \in]0,1[$ such that (5) is satisfied.

Proof. Rewriting the left-hand side of the equation as $x \log w' + \log c'$, where

$$w' = \frac{p(1-q)}{q(1-p)} \quad \text{and} \quad c' = \frac{1-p}{1-q},$$

it follows that $p = q$ is equivalent to $w' = c' = 1$. By definition of c', $p < q$ is equivalent to $c' > 1$ and, as $w'c' = p/q$, this is also equivalent to $c' < 1/w'$. Analogously, it follows that $p > q$ is equivalent to $0 < 1/w' < c' < 1$. By defining $w = \log w'$ and $c = \log c'$ and taking logarithms in the equalities and inequalities, we conclude that $p, q \in]0,1[$ is equivalent to $w \in \mathbb{R}$ and $c = -bw$ with $b \in]0,1[$. □

Proof of Theorem 3. Clearly, the theorem holds for $n = 1$. Suppose, therefore, that $n \geq 2$. According to Corollary 2, $\text{Edim}(\mathcal{N}_0) \leq n+1$. Thus, by Lemma 1 it suffices to show that $\text{VCdim}(\mathcal{N}_0) \geq n+1$. Let e_i denote the vector with a one in the ith position and zeros elsewhere. Further, let $\bar{1}$ be the vector with a 1 in each position. We show that the set of $n+1$ vectors $e_1, \ldots, e_n, \bar{1}$ is shattered by the class $\mathcal{C}_{\mathcal{N}_0}$ of concepts induced by $\mathcal{N}_0$, consisting of the functions of the form

$$\text{sign}\left(\log \frac{P(x)}{Q(x)}\right) \;=\; \text{sign}\left(\sum_{i=1}^{n} x_i \log \frac{p_i}{q_i} + (1-x_i) \log \frac{1-p_i}{1-q_i}\right),$$

where $p_i, q_i \in]0,1[$, for $i \in \{1, \ldots, n\}$. By Lemma 2, the functions in $\mathcal{C}_{\mathcal{N}_0}$ can be written as

$$\text{sign}(w^\top (x - b)),$$

where $w \in \mathbb{R}^n$ and $b \in]0,1[^n$.

It is not difficult to see that homogeneous halfspaces, that is, where $b = (0, \ldots, 0)$, can dichotomize the set $\{e_1, \ldots, e_n, \bar{1}\}$ in all possible ways, except for the two cases to separate $\bar{1}$ from $e_1, \ldots, e_n$. To accomplish these two dichotomies we define $b = (3/4) \cdot \bar{1}$ and $w = \pm\bar{1}$. Then, by the assumption that $n \geq 2$, we have for $i = 1, \ldots, n$,

$$w^\top (e_i - b) \;=\; \pm(1 - 3n/4) \;\lesseqgtr\; 0 \quad \text{and} \quad w^\top (\bar{1} - b) \;=\; \pm(n - 3n/4) \;\gtreqless\; 0.$$

□

A further type of Bayesian network for which we derive the exact dimension has some kind of bipartite graph underlying where one set of nodes serves as the set of parents for all nodes in the other set.

Theorem 4. *For $k \geq 0$, let $\mathcal{N}'_k$ denote the unconstrained network with $P_i = \emptyset$ for $i = 1, \ldots, k$ and $P_i = \{1, \ldots, k\}$ for $i = k+1, \ldots, n$. Then,*

$$\text{Edim}(\mathcal{N}'_k) \;=\; 2^k(n-k+1).$$

Proof. For the upper bound, we apply Lemma 1 and Theorem 1 using the fact that

$$\bigcup_{i=1}^{n} 2^{P_i \cup \{x_i\}} \;=\; \bigcup_{i=k+1}^{n} \{J_i \cup \{i\} \mid J_i \subseteq \{1,\ldots,k\}\} \cup \{J \mid J \subseteq \{1,\ldots,k\}\}.$$

To obtain the lower bound, let $M \subseteq \{0,1\}^{n-k}$ denote the set from the proof of Theorem 3 for the corresponding network $\mathcal{N}_0$ with $n-k$ nodes. We show that the set $S = \{0,1\}^k \times M \subseteq \{0,1\}^n$ is shattered by $\mathcal{N}'_k$. Note that S has the claimed cardinality since $|M| = n-k+1$.

Let (S^-, S^+) be a dichotomy of S (that is, where $S^- \cup S^+ = S$ and $S^- \cap S^+ = \emptyset$). Given a natural number $j \in \{0,\ldots,2^k-1\}$, we use $\mathrm{bin}(j)$ to denote the binary representation of j using k bits. Then, let (M_j^-, M_j^+) be the dichotomy of M defined by

$$M_j^+ \;=\; \{v \in M \mid \mathrm{bin}(j)v \in S^+\}.$$

Here, $\mathrm{bin}(j)v$ refers to the concatenation of the k bits of $\mathrm{bin}(j)$ and the $n-k$ bits of v. According to Theorem 3, for each dichotomy (M_j^-, M_j^+) there exist parameter values p_i^j, q_i^j, where $1 \leq i \leq n-k$, such that $\mathcal{N}_0$ with these parameter settings induces this dichotomy on M. In the network $\mathcal{N}'_k$, we specify the parameters as follows. For $i = 1,\ldots,k$, let

$$p_i \;=\; q_i \;=\; 1/2,$$

and for $i = k+1,\ldots,n$ and each $j \in \{0,\ldots,2^k-1\}$ define

$$\begin{aligned} p_{i,\mathrm{bin}(j)} &= p_{i-k}^j, \\ q_{i,\mathrm{bin}(j)} &= q_{i-k}^j. \end{aligned}$$

Obviously, the concept thus defined by $\mathcal{N}'_k$ outputs -1 for elements of S^- and 1 for elements of S^+. Since every dichotomy of S can be implemented in this way, S is shattered by $\mathcal{N}'_k$. □

4.2 General Lower Bounds

In Section 4.2.1 we shall establish lower bounds on $\mathrm{Edim}(\mathcal{N})$ for unconstrained Bayesian networks and in Section 4.2.2 for networks with a reduced parameter collection. These results are obtained by providing embeddings of concept classes, as introduced in Section 2.1, into these networks. Since $\mathrm{VCdim}(\mathcal{C}) \leq \mathrm{VCdim}(\mathcal{C}')$ if $\mathcal{C} \leq \mathcal{C}'$, a lower bound on $\mathrm{VCdim}(\mathcal{C}')$ follows immediately from classes satisfying $\mathcal{C} \leq \mathcal{C}'$ if the VC dimension of $\mathcal{C}$ is known or easy to determine. We first define concept classes that will suit this purpose.

Definition 7. *Let $\mathcal{N}$ be an arbitrary Bayesian network. For every $i \in \{1, \ldots, n\}$, let $\mathcal{F}_i$ be a family of ± 1-valued functions on the domain $\{0,1\}^{m_i}$ and let $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. Then $\mathcal{C}_{\mathcal{N},\mathcal{F}}$ is the concept class over the domain $\{0,1\}^n \backslash \{(0,\ldots,0)\}$ consisting of all functions of the form*

$$L_{\mathcal{N},f} \quad = \quad [(x_n, f_n), \ldots, (x_1, f_1)],$$

where $f = (f_1, \ldots, f_n) \in \mathcal{F}$. The right-hand side of this equation is to be understood as a decision list, where $L_{\mathcal{N},f}(x)$ for $x \neq (0,\ldots,0)$ is determined as follows:

1. *Find the largest i such that $x_i = 1$.*
2. *Apply f_i to the projection of x to the parent-variables of x_i and output the result.*

The VC dimension of $\mathcal{C}_{\mathcal{N},\mathcal{F}}$ can be directly obtained from the VC dimensions of the classes $\mathcal{F}_i$.

Lemma 3. *Let $\mathcal{N}$ be an arbitrary Bayesian network. Then,*

$$\mathrm{VCdim}(\mathcal{C}_{\mathcal{N},\mathcal{F}}) \quad = \quad \sum_{i=1}^{n} \mathrm{VCdim}(\mathcal{F}_i).$$

Proof. We show that $\mathrm{VCdim}(\mathcal{C}_{\mathcal{N},\mathcal{F}}) \geq \sum_{i=1}^{n} \mathrm{VCdim}(\mathcal{F}_i)$; the proof for the other direction is similar. For every i, we embed the vectors from $\{0,1\}^{m_i}$ into $\{0,1\}^n$ according to $\tau_i(a) = (a', 1, 0, \ldots, 0)$, where $a' \in \{0,1\}^{i-1}$ is chosen such that its projection to the parent-variables of x_i is equal to a and the remaining components are set to 0. Note that $\tau_i(a)$ is absorbed by item (x_i, f_i) of the decision list $L_{\mathcal{N},f}$. It is easy to see that the following holds: If, for $i = 1, \ldots, n$, S_i is a set that is shattered by $\mathcal{F}_i$, then $\cup_{i=1}^{n} \tau_i(S_i)$ is shattered by $\mathcal{C}_{\mathcal{N},\mathcal{F}}$. Thus, $\mathrm{VCdim}(\mathcal{C}_{\mathcal{N},\mathcal{F}}) \geq \sum_{i=1}^{n} \mathrm{VCdim}(\mathcal{F}_i)$. □

The preceding definition and lemma are valid for unconstrained as well as constrained networks as they make use only of the graph underlying the network and do not refer to the values of the parameters. This will be important in the applications that follow.

4.2.1 Lower Bounds for Unconstrained Bayesian Networks

The next theorem is the main step in deriving for an arbitrary unconstrained network $\mathcal{N}$ a lower bound on $\mathrm{Edim}(\mathcal{N})$. It is based on the idea of embedding one of the concept classes $\mathcal{C}_{\mathcal{N},\mathcal{F}}$ defined above into $\mathcal{C}_{\mathcal{N}}$.

Theorem 5. *Let $\mathcal{N}$ be an unconstrained Bayesian network and let $\mathcal{F}_i^*$ denote the set of all ± 1-valued functions on domain $\{0,1\}^{m_i}$. Further, let $\mathcal{F}^* = \mathcal{F}_1^* \times \cdots \times \mathcal{F}_n^*$. Then, $\mathcal{C}_{\mathcal{N},\mathcal{F}^*} \leq \mathcal{C}_{\mathcal{N}}$.*

Proof. We have to show that, for every $f = (f_1, \ldots, f_n)$, we can find a pair (P, Q) of distributions from $\mathcal{D}_{\mathcal{N}}$ such that, for every $x \in \{0,1\}^n$, $L_{\mathcal{N},f}(x) = \operatorname{sign}(\log(P(x)/Q(x)))$. To this end, we define the parameters for the distributions P and Q as

$$p_{i,\alpha} = \begin{cases} 2^{-2^{i-1}n}/2 & \text{if } f_i(\alpha) = -1, \\ 1/2 & \text{if } f_i(\alpha) = +1, \end{cases} \quad \text{and} \quad q_{i,\alpha} = \begin{cases} 1/2 & \text{if } f_i(\alpha) = -1, \\ 2^{-2^{i-1}n}/2 & \text{if } f_i(\alpha) = +1. \end{cases}$$

An easy calculation now shows that

$$\log\left(\frac{p_{i,\alpha}}{q_{i,\alpha}}\right) = f_i(\alpha)2^{i-1}n \quad \text{and} \quad \left|\log\frac{1-p_{i,\alpha}}{1-q_{i,\alpha}}\right| < 1. \tag{6}$$

Fix some arbitrary $x \in \{0,1\}^n \setminus \{(0,\ldots,0)\}$. Choose i_* maximal such that $x_{i_*} = 1$ and let α_* denote the projection of x to the parent-variables of x_{i_*}. Then, $L_{\mathcal{N},f}(x) = f_{i_*}(\alpha_*)$. Thus, $L_{\mathcal{N},f}(x) = \operatorname{sign}(\log(P(x)/Q(x)))$ would follow immediately from

$$\operatorname{sign}\left(\log\frac{P(x)}{Q(x)}\right) \;=\; \operatorname{sign}\left(\log\frac{p_{i_*,\alpha_*}}{q_{i_*,\alpha_*}}\right) \;=\; f_{i_*}(\alpha_*). \tag{7}$$

The second equation in (7) is evident from the equality established in (6). As for the first equation in (7), we argue as follows. By the choice of i_*, we have $x_i = 0$ for every $i > i_*$. Expanding P and Q as given in (3), we obtain

$$\begin{aligned} \log\frac{P(x)}{Q(x)} \;=\; & \log\frac{p_{i_*,\alpha_*}}{q_{i_*,\alpha_*}} + \sum_{i=1}^{i_*-1}\left(\sum_{\alpha\in\{0,1\}^{m_i}} x_i M_{i,\alpha}(x)\log\frac{p_{i,\alpha}}{q_{i,\alpha}}\right) \\ & + \sum_{i\in I}\left(\sum_{\alpha\in\{0,1\}^{m_i}} (1-x_i) M_{i,\alpha}(x)\log\frac{1-p_{i,\alpha}}{1-q_{i,\alpha}}\right), \end{aligned}$$

where $I = \{1,\ldots,n\} \setminus \{i_*\}$. Employing the inequality from (6), it follows that the sign of the right-hand side of this equation is determined by $\log(p_{i_*,\alpha_*}/q_{i_*,\alpha_*})$ since this term is of absolute value $2^{i_*-1}n$ and

$$2^{i_*-1}n - \sum_{j=1}^{i_*-1}(2^{j-1}n) - (n-1) \;\geq\; 1. \tag{8}$$

This concludes the proof. □

Using the lower bound obtained from Theorem 5 combined with Lemma 3 and the upper bound provided by Theorem 1, we have a result that is tight up to a factor of 2.

Corollary 3. *Every unconstrained Bayesian network $\mathcal{N}$ satisfies*

$$\sum_{i=1}^{n} 2^{m_i} \;\leq\; \mathrm{Edim}(\mathcal{N}) \;\leq\; \left|\bigcup_{i=1}^{n} 2^{P_i \cup \{i\}}\right| \;\leq\; 2 \cdot \sum_{i=1}^{n} 2^{m_i}.$$

Bounds for the kth-order Markov chain that are optimal up to an additive constant of 1 emerge from the lower bound due to Theorem 5 with Lemma 3 and the upper bound stated in Corollary 2.

Corollary 4. *Let $\mathcal{N}_k$ denote the Bayesian network from Example 1. Then,*

$$(n-k+1)2^k - 1 \;\leq\; \mathrm{Edim}(\mathcal{N}_k) \;\leq\; (n-k+1)2^k.$$

4.2.2 Lower Bounds for Bayesian Networks with a Reduced Parameter Collection

We now show how to obtain bounds for networks with a reduced parameter collection. Similarly as in Section 4.2.1, the major step consists in providing embeddings into these networks. The main result is based on techniques developed for Theorem 5.

Theorem 6. *Let $\mathcal{N}^R$ be a Bayesian network with a reduced parameter collection $(p_{i,c})_{1\leq i\leq n, 1\leq c\leq d_i}$ in the sense of Definition 3. Let $\mathcal{F}_i^{R_i}$ denote the set of all ± 1-valued functions on the domain $\{0,1\}^{m_i}$ that depend on $\alpha \in \{0,1\}^{m_i}$ only through $R_i(\alpha)$. In other words, $f \in \mathcal{F}_i^{R_i}$ holds if and only if there exists a ± 1-valued function g on domain $\{1,\ldots,d_i\}$ such that $f(\alpha) = g(R_i(\alpha))$ for every $\alpha \in \{0,1\}^{m_i}$. Finally, let $\mathcal{F}^R = \mathcal{F}_1^{R_1} \times \cdots \times \mathcal{F}_n^{R_n}$. Then, $\mathcal{C}_{\mathcal{N}^R, \mathcal{F}^R} \leq \mathcal{C}_{\mathcal{N}^R}$.*

Proof. We focus on the differences to the proof of Theorem 5. First, the decision list $L_{\mathcal{N}^R, f}$ uses a function $f = (f_1, \ldots, f_n)$ of the form $f_i(x) = g_i(R_i(x))$ for some function $g_i : \{1,\ldots,d_i\} \to \{-1,1\}$. Second, the distributions P, Q that satisfy $L_{\mathcal{N},f}(x) = \mathrm{sign}(\log(P(x)/Q(x)))$ for every $x \in \{0,1\}^n$ have to be defined over the reduced parameter collection as given in equation (4). An appropriate choice is

$$p_{i,c} = \begin{cases} 2^{-2^{i-1}n}/2 & \text{if } g_i(c) = -1, \\ 1/2 & \text{if } g_i(c) = 1, \end{cases} \quad \text{and} \quad q_{i,c} = \begin{cases} 1/2 & \text{if } g_i(c) = -1, \\ 2^{-2^{i-1}n}/2 & \text{if } g_i(c) = 1. \end{cases}$$

The rest of the proof is completely analogous to the proof of Theorem 5. □

Theorem 5 can be viewed as a special case of Theorem 6 since every unconstrained network can be considered as a network with a reduced parameter collection where the functions R_i are 1-1. However, there are differences arising from the notation of the network parameters that have been taken into account by the above proof.

Applying the lower bound of Theorem 6 in combination with Lemma 3 and the upper bound of Theorem 2, we once more have bounds that are optimal up to the factor 2.

Corollary 5. *Let $\mathcal{N}^R$ be a Bayesian network with a reduced parameter collection $(p_{i,c})_{1\leq i\leq n, 1\leq c\leq d_i}$ in the sense of Definition 3. Then,*

$$\sum_{i=1}^{n} d_i \;\leq\; \mathrm{Edim}(\mathcal{N}^R) \;\leq\; 2\cdot\sum_{i=1}^{n} d_i.$$

4.2.3 Lower Bounds for Logistic Autoregressive Networks

The following result is not obtained by embedding a concept class into a logistic autoregressive Bayesian network. However, we apply a similar technique as developed in Sections 4.2.1 and 4.2.2 to derive a bound using the VC dimension by directly showing that these networks can shatter sets of the claimed size.

Theorem 7. *Let $\mathcal{N}_\sigma$ denote the logistic autoregressive Bayesian network from Definition 4. Then,*

$$\mathrm{Edim}(\mathcal{N}_\sigma) \;\;\geq\;\; n(n-1)/2.$$

Proof. We show that the following set S is shattered by the concept class $\mathcal{C}_{\mathcal{N}_\sigma}$. Then the statement follows from Lemma 1.

For $i = 2,\ldots,n$ and $c = 1,\ldots,i-1$, let $\alpha_{i,c} \in \{0,1\}^{i-1}$ be the pattern with bit 1 in position c and zeros elsewhere. Then, for every pair (i,c), where $i \in \{2,\ldots,n\}$ and $c \in \{1,\ldots,i-1\}$, let $s^{(i,c)} \in \{0,1\}^n$ be the vector that has bit 1 in coordinate i, bit-pattern $\alpha_{i,c}$ in the coordinates $1,\ldots,i-1$, and zeros in the remaining positions. The set

$$S \;\;=\;\; \{s^{(i,c)} \mid i = 2,\ldots,n \text{ and } c = 1,\ldots,i-1\}$$

has $n(n-1)/2$ elements.

To show that S is shattered, let (S^-, S^+) be some arbitrary dichotomy of S. We claim that there exists a pair (P,Q) of distributions from $\mathcal{D}_{\mathcal{N}_\sigma}$ such that for every $s^{(i,c)}$, $\mathrm{sign}(\log(P(s^{(i,c)})/Q(s^{(i,c)}))) = 1$ if and only if $s^{(i,c)} \in S^+$. Assume that the parameters $p_{i,\alpha}$ and $q_{i,\alpha}$ for the distributions P and Q, respectively, satisfy

$$p_{i,\alpha} \;\;=\;\; \begin{cases} 1/2 & \text{if } \alpha = \alpha_{i,c} \text{ and } s^{(i,c)} \in S^+, \\ 2^{-2^{i-1}n}/2 & \text{otherwise,} \end{cases}$$

and

$$q_{i,\alpha} \;\;=\;\; \begin{cases} 2^{-2^{i-1}n}/2 & \text{if } \alpha = \alpha_{i,c} \text{ and } s^{(i,c)} \in S^+, \\ 1/2 & \text{otherwise.} \end{cases}$$

Similarly as in the proof of Theorem 5, we have

$$\left|\log\left(\frac{p_{i,\alpha}}{q_{i,\alpha}}\right)\right| = 2^{i-1}n \quad\text{and}\quad \left|\log\frac{1-p_{i,\alpha}}{1-q_{i,\alpha}}\right| < 1. \tag{9}$$

The expansion of P and Q yields for every $s^{(i,c)} \in S$,

$$\log \frac{P(s^{(i,c)})}{Q(s^{(i,c)})} = \log \frac{p_{i,\alpha_{i,c}}}{q_{i,\alpha_{i,c}}} + \sum_{j=1}^{i-1} \left(\sum_{\alpha \in \{0,1\}^{j-1}} s_j^{(i,c)} M_{j,\alpha}(s^{(i,c)}) \log \frac{p_{j,\alpha}}{q_{j,\alpha}} \right) + \sum_{j \in I} \left(\sum_{\alpha \in \{0,1\}^{j-1}} (1 - s_j^{(i,c)}) M_{j,\alpha}(s^{(i,c)}) \log \frac{1 - p_{j,\alpha}}{1 - q_{j,\alpha}} \right),$$

where $I = \{1, \ldots, n\} \backslash \{i\}$. In analogy to inequality (8) in the proof of Theorem 5, it follows from (9) that the sign of $\log(P(s^{(i,c)})/Q(s^{(i,c)}))$ is equal to the sign of $\log(p_{i,\alpha_{i,c}}/q_{i,\alpha_{i,c}})$. By the definition of $p_{i,\alpha_{i,c}}$ and $q_{i,\alpha_{i,c}}$, the sign of $\log(p_{i,\alpha_{i,c}}/q_{i,\alpha_{i,c}})$ is positive if and only if $s^{(i,c)} \in S^+$.

It remains to show that the parameters of the distributions P and Q can be given as required by Definition 4, that is, in the form $p_{i,\alpha} = \sigma(\sum_{j=1}^{i-1} w_{i,j}\alpha_j)$ with $w_{i,j} \in \mathbb{R}$, and similarly for $q_{i,\alpha}$. This now immediately follows from the fact that $\sigma(\mathbb{R}) =]0, 1[$. □

5 Lower Bounds via Embeddings of Parity Functions

The lower bounds obtained in Section 4 rely on arguments based on the VC dimension of the respective concept class. In particular, a quadratic lower bound for the logistic autoregressive network has been established. In the following, we introduce a different technique leading to the lower bound $2^{\Omega(n)}$ for a variant of this network. For the time being, it seems possible to obtain an exponential bound for these slightly modified networks only, which are given by the following definition.

Definition 8. *The* modified logistic autoregressive Bayesian network $\mathcal{N}'_\sigma$ *is the fully connected Bayesian network with nodes* $0, 1, \ldots, n+1$ *and the constraints on the parameter collection defined as*

$$\forall i = 0, \ldots, n, \exists (w_{i,j})_{0 \le j \le i-1} \in \mathbb{R}^i, \forall \alpha \in \{0,1\}^i : p_{i,\alpha} = \sigma \left(\sum_{j=0}^{i-1} w_{i,j}\alpha_j \right)$$

and

$$\exists (w_i)_{0 \le i \le n}, \forall \alpha \in \{0,1\}^{n+1} : p_{n+1,\alpha} = \sigma \left(\sum_{i=0}^{n} w_i \sigma \left(\sum_{j=0}^{i-1} w_{i,j}\alpha_j \right) \right).$$

Obviously, the network $\mathcal{N}'_\sigma$ *is completely described by the parameter collections* $(w_{i,j})_{0 \le i \le n, 0 \le j \le i-1}$ *and* $(w_i)_{0 \le i \le n}$.

The crucial difference between $\mathcal{N}'_\sigma$ and $\mathcal{N}_\sigma$ is the node $n+1$ whose sigmoidal function receives the outputs of the other sigmoidal functions as input. Roughly speaking, $\mathcal{N}_\sigma$ is a single-layer network whereas $\mathcal{N}'_\sigma$ has an extra node at a second layer.

To obtain the bound, we provide an embedding of the concept class of parity functions. The following theorem motivates this construction by showing that it is impossible to obtain an exponential lower bound for $\mathrm{Edim}(\mathcal{N}_\sigma)$ nor for $\mathrm{Edim}(\mathcal{N}'_\sigma)$ using the VC dimension argument, as these networks have VC dimensions that are polynomial in n.

Theorem 8. *The logistic autoregressive Bayesian network $\mathcal{N}_\sigma$ from Definition 4 and the modified logistic autoregressive Bayesian network $\mathcal{N}'_\sigma$ from Definition 8 have a VC dimension that is bounded by $O(n^6)$.*

Proof. Consider first the logistic autoregressive Bayesian network. We show that the concept class induced by $\mathcal{N}_\sigma$ can be computed by a specific type of feedforward neural network. Then, we apply a known bound on the VC dimension of these networks.

The neural networks for the concepts in $\mathcal{C}_{\mathcal{N}_\sigma}$ consist of sigmoidal units, product units, and units computing second-order polynomials. A sigmoidal unit computes functions of the form $\sigma(w^\top x - t)$, where $x \in \mathbb{R}^k$ is the input vector and $w \in \mathbb{R}^k, t \in \mathbb{R}$ are parameters. A product unit computes $\Pi_{i=1}^k x_i^{w_i}$.

The value of $p_{i,\alpha}$ can be calculated by a sigmoidal unit as $p_{i,\alpha} = \sigma(\sum_{j=1}^{i-1} w_{i,j}\alpha_j)$ with α as input and parameters $w_{i,1}, \ldots, w_{i,i-1}$. Regarding the factors $p_{i,\alpha}^{x_i}(1 - p_{i,\alpha})^{(1-x_i)}$, we observe that

$$\begin{aligned} p_{i,\alpha}^{x_i}(1 - p_{i,\alpha})^{(1-x_i)} &= p_{i,\alpha}x_i + (1 - p_{i,\alpha})(1 - x_i) \\ &= 2p_{i,\alpha}x_i - x_i - p_{i,\alpha} + 1, \end{aligned}$$

where the first equation is valid because $x_i \in \{0, 1\}$. Thus, the value of $p_{i,\alpha}^{x_i}(1 - p_{i,\alpha})^{(1-x_i)}$ is given by a second-order polynomial. Similarly, the value of $q_{i,\alpha}^{x_i}(1 - q_{i,\alpha})^{(1-x_i)}$ can also be determined using sigmoidal units and polynomial units of order 2. Finally, the output value of the network is obtained by comparing $P(x)/Q(x)$ with the constant threshold 1. We calculate $P(x)/Q(x)$ using a product unit

$$y_1 \cdots y_n z_1^{-1} \cdots z_n^{-1},$$

with input variables y_i and z_i that receive the value of $p_{i,\alpha}^{x_i}(1 - p_{i,\alpha})^{(1-x_i)}$ and $q_{i,\alpha}^{x_i}(1 - q_{i,\alpha})^{(1-x_i)}$ computed by the second-order units, respectively.

This network has $O(n^2)$ parameters and $O(n)$ computation nodes, each of which is a sigmoidal unit, a second-order unit, or a product unit. Theorem 2 of Schmitt (2002) shows that every such network with W parameters and k computation nodes, which are sigmoidal and product units, has VC dimension

$O(W^2k^2)$. A close inspection of the proof of this result reveals that it also includes polynomials of degree 2 as computational units (see also Lemma 4 in Schmitt, 2002). Thus, we obtain the claimed bound $O(n^6)$ for the logistic autoregressive Bayesian network $\mathcal{N}_\sigma$.

For the modified logistic autoregressive network we have only to take one additional sigmoidal unit into account. Thus, the bound for this network follows now immediately. □

In the previous result we were interested in the asymptotic behavior of the VC dimension, showing that it is not exponential. Using the techniques provided in Schmitt (2002) mentioned in the above proof, it is also possible to obtain constant factors for these bounds.

We now provide the main result of this section. Its proof employs the concept class PARITY_n defined in Section 2.3.

Theorem 9. *Let $\mathcal{N}'_\sigma$ denote the modified logistic autoregressive Bayesian network with $n+2$ nodes and assume that n is a multiple of 4. Then,* $\mathrm{PARITY}_{n/2} \leq \mathcal{N}'_\sigma$.

Proof. The mapping

$$\{0,1\}^{n/2} \ni x = (x_1, \ldots, x_{n/2}) \mapsto (\overbrace{1, x_1, \ldots, x_{n/2}, 1, \ldots, 1}^{\alpha}, 1) = x' \in \{0,1\}^{n+2} \tag{10}$$

assigns to every element of $\{0,1\}^{n/2}$ uniquely some element in $\{0,1\}^{n+2}$. Note that α, as indicated in (10), equals the bit-pattern of the parent-variables of x'_{n+2} (which are actually all other variables). We claim that the following holds. For every $a \in \{0,1\}^{n/2}$, there exists a pair (P,Q) of distributions from $\mathcal{D}_{\mathcal{N}'_\sigma}$ such that for every $x \in \{0,1\}^{n/2}$,

$$(-1)^{a^\top x} \;=\; \mathrm{sign}\left(\log \frac{P(x')}{Q(x')}\right). \tag{11}$$

Clearly, the theorem follows once the claim is settled. The proof of the claim makes use of the following facts:

Fact 1 For every $a \in \{0,1\}^{n/2}$, function $(-1)^{a^\top x}$ can be computed by a two-layer threshold circuit with $n/2$ threshold units at the first layer and one threshold unit as output node at the second layer.

Fact 2 Each two-layer threshold circuit C can be simulated by a two-layer sigmoidal circuit C' with the same number of units and the following output convention: $C(x) = 1 \implies C'(x) \geq 2/3$ and $C(x) = 0 \implies C'(x) \leq 1/3$.

Fact 3 Network $\mathcal{N}'_\sigma$ contains as a sub-network a two-layer sigmoidal circuit C' with $n/2$ input nodes, $n/2$ sigmoidal units at the first layer, and one sigmoidal unit at the second layer.

The parity function is a symmetric Boolean function, that is, a function $f : \{0,1\}^k \to \{0,1\}$ that is described by a set $M \subseteq \{0,\ldots,k\}$ such that $f(x) = 1$ if and only if $\sum_{i=1}^k x_i \in M$. Thus, Fact 1 is implied by Proposition 2.1 of Hajnal et al. (1993) which shows that every symmetric Boolean function can be computed by a circuit of this kind.

Fact 2 follows from the capability of the sigmoidal function σ to approximate any Boolean threshold function arbitrarily close. This can be done by multiplying all weights and the threshold with a sufficiently large number.

To establish Fact 3, we refer to Definition 8 and proceed as follows: We would like the term $p_{n+1,\alpha}$ to satisfy $p_{n+1,\alpha} = C'(\alpha_1,\ldots,\alpha_{n/2})$, where C' denotes an arbitrary two-layer sigmoidal circuit as described in Fact 3. To this end, we set $w_{i,j} = 0$ if $1 \leq i \leq n/2$ or if $i,j \geq n/2+1$. Further, we let $w_i = 0$ if $1 \leq i \leq n/2$. The parameters that have been set to zero are referred to as "redundant" parameters in what follows. Recall from (10) that $\alpha_0 = \alpha_{n/2+1} = \cdots = \alpha_n = 1$. From these settings and from $\sigma(0) = 1/2$, we obtain

$$p_{n+1,\alpha} = \sigma\left(\frac{1}{2}w_0 + \sum_{i=n/2+1}^{n} w_i \sigma\left(w_{i,0} + \sum_{j=1}^{n/2} w_{i,j}\alpha_j\right)\right).$$

Indeed, this is the output of a two-layer sigmoidal circuit C' on the input $(\alpha_1,\ldots,\alpha_{n/2})$.

We are now in the position to describe the choice of distributions P and Q. Let C' be the sigmoidal circuit that computes $(-1)^{a^\top x}$ for some fixed $a \in \{0,1\}^{n/2}$ according to Facts 1 and 2. Let P be the distribution obtained by setting the redundant parameters to zero (as described above) and the remaining parameters as in C'. Thus, $p_{n+1,\alpha} = C'(\alpha_1,\ldots,\alpha_{n/2})$. Let Q be the distribution with the same parameters as P except for replacing w_i by $-w_i$. Thus, by symmetry of σ, $q_{n+1,\alpha} = 1 - C'(\alpha_1,\ldots,\alpha_{n/2})$. Since $x'_{n+1} = 1$ and since all but one factor in $P(x')/Q(x')$ cancel each other, we arrive at

$$\frac{P(x')}{Q(x')} \;=\; \frac{p_{n+1,\alpha}}{q_{n+1,\alpha}} \;=\; \frac{C'(\alpha_1,\ldots,\alpha_{n/2})}{1 - C'(\alpha_1,\ldots,\alpha_{n/2})}.$$

As C' computes $(-1)^{a^\top x}$, with the output convention from Fact 2 we have that $P(x')/Q(x') \geq 2$ if $(-1)^{a^\top x} = 1$, and $P(x')/Q(x') \leq 1/2$ otherwise. This implies claim (11) and concludes the proof. □

Combining Theorem 9 with Corollary 1, we obtain the exponential lower bound for the modified logistic autoregressive Bayesian network.

Corollary 6. *Let $\mathcal{N}'_\sigma$ denote the modified logistic autoregressive Bayesian network. Then,* $\mathrm{Edim}(\mathcal{N}'_\sigma) \geq 2^{n/4}$.

By a more detailed analysis it can be shown that Theorem 9 holds even if we restrict the values in the parameter collection of $\mathcal{N}'_\sigma$ to integers that can be represented using $O(\log n)$ bits. We mentioned in the introduction that a large lower bound on $\mathrm{Edim}(\mathcal{C})$ rules out the possibility of a large margin classifier. For the class PARITY_n, this can be made more precise. Forster and Simon (2002) have shown that every linear arrangement for PARITY_n has an average geometric margin of at most $2^{-n/2}$. Thus there can be no linear arrangement with an average margin exceeding $2^{-n/4}$ for $\mathcal{C}_{\mathcal{N}'_\sigma}$ even if we restrict the weight parameters in $\mathcal{N}'_\sigma$ to logarithmically bounded integers.

6 Conclusions and Open Problems

Bayesian networks have become one of the heavily studied and widely used probabilistic techniques for pattern recognition and statistical inference. One line of inquiry into Bayesian networks pursues the idea of combining them with kernel methods so that one can take advantage of both. Kernel methods employ the principle of mapping the input vectors to some higher-dimensional space where then inner product operations are performed implicitly. The major motivation for our work was to reveal more about this inner product space. In particular, we were interested to know whether the kernel operations in the higher-dimensional space can be implemented as standard dot products. With the results we have gained insight into the nature of the inner product space in terms of bounds on its dimensionality. As the main results, we have established tight bounds on the Euclidean dimension of spaces in which two-label classifications of Bayesian networks with binary nodes can be implemented.

As one of the tools for deriving lower bounds we have employed the VC dimension. Bounds on the VC dimension of concept classes abound. Only for a few classes the exact values are known. Surprisingly, our investigation of the dimensionality of embeddings lead to some exact values of the Euclidean dimension and the VC dimension for nontrivial Bayesian networks. Moreover, the fact that in the main results Euclidean dimension and VC dimension are closely tied together has implications for the complexity of model selection, that is, for the amount of information required for choosing a Bayesian network that performs well on unseen data. For a setting where this magnitude can be measured in terms of the VC dimension, the tight bounds for the embeddings of Bayesian networks show that the amount of data required for learning can be estimated using the Euclidean dimension. For the logistic autoregressive network, this is an open issue. Whether the VC dimension is also useful in tightly bounding the Euclidean dimension of these networks remains to be shown. For the modified

version of this model, our results suggest that different approaches might be more successful.

The results raise some further open questions. First, since we considered only networks with binary nodes, analogous questions regarding Bayesian networks with multiple-valued or even continuous-valued nodes are certainly of interest. Another generalization of Bayesian networks are those with hidden variables which have also been out of the scope of this work. Further, with regard to logistic autoregressive Bayesian networks, we were able to obtain an exponential lower bound only for a variant of them. For the unmodified network, such a bound has yet to be found. Finally, the questions we studied here are certainly relevant not only for Bayesian networks but also for other popular classes of distributions or densities. Those from the exponential family look like a good thing to start with.

Acknowledgments

We thank the anonymous reviewers for the many helpful and comprehensive comments.

This work was supported in part by the IST Programme of the European Community under the PASCAL Network of Excellence, IST-2002-506778, and by the Deutsche Forschungsgemeinschaft (DFG), grant SI 498/7-1.

References

Altun, Y., Tsochantaridis, I., and Hofmann, T. (2003). Hidden Markov support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, pages 3–10. AAAI Press, Menlo Park, CA.

Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.

Arriaga, R. I. and Vempala, S. (1999). An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 616–623. IEEE Computer Society Press, Los Alamitos, CA.

Balcan, M.-F., Blum, A., and Vempala, S. (2004). On kernels, margins, and low-dimensional mappings. In Ben-David, S., Case, J., and Maruoka, A., editors, *Proceedings of the 15th International Conference on Algorithmic Learning Theory ALT 2004*, volume 3244 of *Lecture Notes in Artificial Intelligence*, pages 194–205. Springer-Verlag, Berlin.

Ben-David, S., Eiron, N., and Simon, H. U. (2002). Limitations of learning via embeddings in Euclidean half-spaces. *Journal of Machine Learning Research*, 3:441–461.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, New York, NY.

Chickering, D. M., Heckerman, D., and Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 80–89. Morgan Kaufmann, San Francisco, CA.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, Berlin.

Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley & Sons, New York, NY.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929.

Forster, J. (2002). A linear lower bound on the unbounded error communication complexity. *Journal of Computer and System Sciences*, 65:612–625.

Forster, J., Krause, M., Lokam, S. V., Mubarakzjanov, R., Schmitt, N., and Simon, H. U. (2001). Relations between communication complexity, linear arrangements, and computational complexity. In Hariharan, R., Mukund, M., and Vinay, V., editors, *Proceedings of the 21st Annual Conference on the Foundations of Software Technology and Theoretical Computer Science*, volume 2245 of *Lecture Notes in Computer Science*, pages 171–182. Springer-Verlag, Berlin.

Forster, J., Schmitt, N., Simon, H. U., and Suttorp, T. (2003). Estimating the optimal margins of embeddings in Euclidean halfspaces. *Machine Learning*, 51:263–281.

Forster, J. and Simon, H. U. (2002). On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. In Cesa-Bianchi, N., Numao, M., and Reischuk, R., editors, *Proceedings of the 13th International Workshop on Algorithmic Learning Theory ALT 2002*, volume 2533 of *Lecture Notes in Artificial Intelligence*, pages 128–138. Springer-Verlag, Berlin.

Frankl, P. and Maehara, H. (1988). The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44:355–362.

Frey, B. J. (1998). *Graphical Models for Machine Learning and Digital Communication*. MIT Press, Cambridge, MA.

Hajnal, A., Maass, W., Pudlák, P., Szegedy, M., and Turán, G. (1993). Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46:129–154.

Jaakkola, T. S. and Haussler, D. (1999a). Exploiting generative models in discriminative classifiers. In Kearns, M. S., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, Cambridge, MA.

Jaakkola, T. S. and Haussler, D. (1999b). Probabilistic kernel regression models. In Heckerman, D. and Whittaker, J., editors, *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics*. Morgan Kaufmann, San Francisco, CA.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipshitz mapping into Hilbert spaces. *Contemporary Mathematics*, 26:189–206.

Kiltz, E. (2003). On the representation of Boolean predicates of the Diffie-Hellman function. In Alt, H. and Habib, M., editors, *Proceedings of 20th International Symposium on Theoretical Aspects of Computer Science*, volume 2607 of *Lecture Notes in Computer Science*, pages 223–233. Springer-Verlag, Berlin.

Kiltz, E. and Simon, H. U. (2003). Complexity theoretic aspects of some cryptographic functions. In Warnow, T. and Zhu, B., editors, *Proceedings of the 9th International Conference on Computing and Combinatorics COCOON 2003*, volume 2697 of *Lecture Notes in Computer Science*, pages 294–303. Springer-Verlag, Berlin.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman and Hall, London.

Nakamura, A., Schmitt, M., Schmitt, N., and Simon, H. U. (2004). Bayesian networks and inner product spaces. In Shawe-Taylor, J. and Singer, Y., editors, *Proceedings of the 17th Annual Conference on Learning Theory COLT 2004*, volume 3120 of *Lecture Notes in Artificial Intelligence*, pages 518–533. Springer-Verlag, Berlin.

Neal, R. M. (1992). Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113.

Oliver, N., Schölkopf, B., and Smola, A. J. (2000). Natural regularization from generative models. In Smola, A. J., Bartlett, P. L., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, pages 51–60. MIT Press, Cambridge, MA.

Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence*, pages 133–136. AAAI Press, Menlo Park, CA.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76.

Saunders, C., Shawe-Taylor, J., and Vinokourov, A. (2003). String kernels, Fisher kernels and finite state automata. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 633–640. MIT Press, Cambridge, MA.

Schmitt, M. (2002). On the complexity of computing and learning with multiplicative neural networks. *Neural Computation*, 14:241–301.

Spiegelhalter, D. J. and Knill-Jones, R. P. (1984). Statistical and knowledge-based approaches to clinical decision support systems. *Journal of the Royal Statistical Society, Series A*, 147:35–77.

Taskar, B., Guestrin, C., and Koller, D. (2004). Max-margin Markov networks. In Thrun, S., Saul, L. K., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, pages 25–32. MIT Press, Cambridge, MA.

Tsuda, K., Akaho, S., Kawanabe, M., and Müller, K.-R. (2004). Asymptotic properties of the Fisher kernel. *Neural Computation*, 16:115–137.

Tsuda, K. and Kawanabe, M. (2002). The leave-one-out kernel. In Dorronsoro, J. R., editor, *Proceedings of the International Conference on Artificial Neural Networks ICANN 2002*, volume 2415 of *Lecture Notes in Computer Science*, pages 727–732. Springer-Verlag, Berlin.

Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S., and Müller, K.-R. (2002). A new discriminative kernel from probabilistic models. *Neural Computation*, 14:2397–2414.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley & Sons, New York, NY.