Bayesian Networks and Inner Product Spaces ⋆

Atsuyoshi Nakamura[1], Michael Schmitt[2],
Niels Schmitt[2], and Hans Ulrich Simon[2]

[1] Graduate School of Engineering, Hokkaido University, Sapporo 060-8628, Japan
atsu@main.eng.hokudai.ac.jp
[2] Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany
{mschmitt,nschmitt,simon}@lmi.ruhr-uni-bochum.de

Abstract. In connection with two-label classification tasks over the Boolean domain, we consider the possibility to combine the key advantages of Bayesian networks and of kernel-based learning systems. This leads us to the basic question whether the class of decision functions induced by a given Bayesian network can be represented within a low-dimensional inner product space. For Bayesian networks with an explicitly given (full or reduced) parameter collection, we show that the "natural" inner product space has the smallest possible dimension up to factor 2 (even up to an additive term 1 in many cases). For a slight modification of the so-called logistic autoregressive Bayesian network with n nodes, we show that every sufficiently expressive inner product space has dimension at least $2^{n/4}$. The main technical contribution of our work consists in uncovering combinatorial and algebraic structures within Bayesian networks such that known techniques for proving lower bounds on the dimension of inner product spaces can be brought into play.

1 Introduction

During the last decade, there has been a lot of interest in learning systems whose hypotheses can be written as inner products in an appropriate feature space, trained with a learning algorithm that performs a kind of empirical or structural risk minimization. The inner product operation is often not carried out explicitly, but reduced to the evaluation of a so-called kernel-function that operates on instances of the original data space, which offers the opportunity to handle high-dimensional feature spaces in an efficient manner. This learning strategy introduced by Vapnik and co-workers [4, 33] 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.

Bayesian networks have a long history in statistics, and in the first half of the 1980s they were introduced to the field of expert systems through work by Pearl [25] and Spiegelhalter and Knill-Jones [29]. They are much different from

⋆ This work has been supported in part by the Deutsche Forschungsgemeinschaft Grant SI 498/7-1.

kernel-based learning systems and offer some complementary advantages. They graphically model conditional independence relationships between random variables. There exist quite elaborated methods for choosing an appropriate network, for performing probabilistic inference (inferring missing data from existing ones), and for solving pattern classification tasks or unsupervised learning problems. Like other probabilistic models, Bayesian networks can be used to represent inhomogeneous training samples with possibly overlapping features and missing data in a uniform manner.

Quite recently, several research groups considered the possibility to combine the key advantages of probabilistic models and kernel-based learning systems. For this purpose, several kernels (like the Fisher-kernel, for instance) were studied extensively [17, 18, 24, 27, 31, 32, 30]. Altun, Tsochantaridis, and Hofmann [1] proposed (and experimented with) a kernel related to the Hidden Markov Model.

In this paper, we focus on two-label classification tasks over the Boolean domain and on probabilistic models that can be represented as Bayesian networks. Intuitively, we aim at finding the "simplest" inner product space that is able to express the class of decision functions (briefly called "concept class" in what follows) induced by a given Bayesian network. We restrict ourselves to Euclidean spaces equipped with the standard scalar product.[3] Furthermore, we use the Euclidean dimension of the space as our measure of simplicity.[4] Our main results are as follows:

1) For Bayesian networks with an explicitly given (full or reduced) parameter collection, the "natural" inner product space (obtained from the probabilistic model by fairly straightforward algebraic manipulations) has the smallest possible dimension up to factor 2 (even up to an additive term 1 in many cases). The (almost) matching lower bounds on the smallest possible dimension are found by analyzing the VC-dimension of the concept class associated with a Bayesian network.

2) We present a quadratic lower bound and the upper bound $O(n^6)$ on the VC-dimension of the concept class associated with the so-called "logistic autoregressive Bayesian network" (also known as "sigmoid belief network")[5], where n denotes the number of nodes.

3) For a slight modification of the logistic autoregressive Bayesian network with $n+2$ nodes, we show that every sufficiently expressive inner product space has dimension at least $2^{n/4}$. The proof of this lower bound proceeds by showing that

[3] This is no loss of generality (except for the infinite-dimensional case) since any finite-dimensional reproducing kernel Hilbert space is isometric with $\mathbb{R}^d$ for some d.

[4] 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 [19, 14, 2], it can be shown that any arrangement with a large margin can be converted into a low-dimensional arrangement. Thus, a large lower bound on the smallest possible dimension rules out the possibility of a large margin classifier.

[5] originally proposed by Mc-Cullagh and Nelder [7] and studied systematically, for instance, by Neal [23] and by Saul, Jaakkola, and Jordan [26]

the concept class induced by such a network contains exponentially many decision functions that are pairwise orthogonal on an exponentially large subdomain. Since the VC-dimension of this concept class has the same order of magnitude, $O(n^6)$, as the original (unmodified) network, VC-dimension considerations would be insufficient to reveal the exponential lower bound.

While (as mentioned above) there exist already some papers that investigate the connection between probabilistic models and inner product spaces, it seems that this work is the first one which addresses explicitly the question of finding a smallest-dimensional sufficiently expressive inner product space. It should be mentioned however that there exist a couple of papers [10, 11, 3, 13, 12, 20, 21] (not concerned with probabilistic models) considering the related question of finding an embedding of a given concept class into a system of half-spaces. The main technical contribution of our work can be seen in uncovering combinatorial and algebraic structures within Bayesian networks such that techniques known from these papers can be brought into play.

2 Preliminaries

In this section, we present formal definitions for the basic notions in this paper. Subsection 2.1 is concerned with notions from learning theory. In Subsection 2.2, we formally introduce Bayesian networks and the distributions and concept classes induced by them. The notion of a linear arrangement for a concept class is presented in Subsection 2.3.

2.1 Concept Classes and VC-Dimension

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 then called a *concept*. A set $S = \{s_1, \ldots, s_m\} \subseteq \mathcal{X}$ of size m is said to be *shattered* by $\mathcal{C}$ if

$$\forall b \in \{-1, +1\}^m, \exists f \in \mathcal{C}, \forall i = 1, \ldots, m : f(s_i) = b_i \enspace .$$

The *VC-dimension* of $\mathcal{C}$ is given by

$$\mathrm{VCdim}(\mathcal{C}) = \sup\{m | \exists S \subseteq \mathcal{X} : |S| = m \text{ and } S \text{ is shattered by } \mathcal{C}\} \enspace .$$

For every $z \in \mathbb{R}$, let $\mathrm{sign}(z) = +1$ if $z \geq 0$ and $\mathrm{sign}(z) = -1$ otherwise. In the context of concept classes, the sign-function is sometimes used for mapping real-valued functions f to ± 1-valued functions $\mathrm{sign} \circ f$.

We write $\mathcal{C} \leq \mathcal{C}'$ for concept classes $\mathcal{C}$ over domain $\mathcal{X}$ and $\mathcal{C}'$ over domain $\mathcal{X}'$ if there exist mappings

$$\mathcal{C} \ni f \mapsto f' \in \mathcal{C}', \mathcal{X} \ni x \mapsto x' \in \mathcal{X}'$$

such that $f(x) = f'(x')$ for every $f \in \mathcal{C}$ and every $x \in \mathcal{X}$. Note that $\mathcal{C} \leq \mathcal{C}'$ implies that $\mathrm{VCdim}(\mathcal{C}) \leq \mathrm{VCdim}(\mathcal{C}')$ because the following holds: if $S \subseteq \mathcal{X}$ is a set of size m that is shattered by $\mathcal{C}$ then $S' = \{s' | s \in S\} \subseteq \mathcal{X}'$ is a set of size m that is shattered by $\mathcal{C}'$.

2.2 Bayesian Networks

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

1. *a directed acyclic graph* $\mathcal{G} = (V, E)$*,*
2. *a collection* $(p_{i,\alpha})_{i \in V, \alpha \in \{0,1\}^{m_i}}$ *of programmable parameters with values in the open intervall* $]0,1[$*, where* m_i *denotes the number of* $j \in V$ *such that* $(j,i) \in E$*,*
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, we say the network is* constrained.

Conventions: We will identify the $n = |V|$ nodes of $\mathcal{N}$ with the numbers from 1 to n and assume that every edge $(j,i) \in E$ satisfies $j < i$ (topological ordering). If $(j,i) \in E$, then j is called a parent of i. P_i denotes the set of parents of node i and $m_i = |P_i|$ denotes the number of parents. $\mathcal{N}$ is said to be *fully connected* if $P_i = 1, \ldots, i-1$ for every node i. We will 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. $M_{i,\alpha}(x)$ denotes the polynomial that indicates whether the parent variables of x_i exhibit bit-pattern α. More formally, $M_{i,\alpha}(x) = \prod_{j \in P_i} x_j^{\alpha_j}$, where $x_j^0 = 1 - x_j$ and $x_j^1 = x_j$.

An unconstrained network with a dense graph has an exponentially growing number of parameters. In a constrained network, the number of parameters can be kept reasonably small even in case of a dense topology. The following two definitions exemplify this approach. Definition 2 contains (as a special case) the networks that were proposed in [5]. (See Example 2 below.) Definition 3 deals with so-called logistic autoregressive Bayesian networks that, given their simplicity, perform surprisingly well on some problems. (See the discussion of these networks in [15].)

Definition 2. *A Bayesian network with a* reduced parameter collection *is a Bayesian network whose constraints can be described as follows. 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*

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

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* $(p_{i,c})_{1 \le i \le n, 1 \le c \le d_i}$.

Definition 3. *The* logistic autoregressive *Bayesian network* $\mathcal{N}_\sigma$ *is the fully connected Bayesian network with the following constraints on the parameter collection:*

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

where $\sigma(y) = 1/(1+e^{-y})$ *denotes 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 the introduction, we mentioned that Bayesian networks graphically model conditional independence relationships. This general idea is captured in the following

Definition 4. *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*

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

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

The polynomial representation of $\log P(x)$ resulting from (1) is called *Chow expansion* in the pattern classification literature [9]. Parameter $p_{i,\alpha}$ represents the conditional probability for $x_i = 1$ given that the parent variables of x_i exhibit bit-pattern α. Formula (1) expresses $P(x)$ as a product of conditional probabilities (chain-expansion).

Example 1 (k-order Markov chain). For $k \ge 0$, $\mathcal{N}_k$ denotes 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 equals $2^k(n-k) + 2^{k-1} + \cdots + 2 + 1 = 2^k(n-k+1) - 1$.

We briefly explain that, for a Bayesian network with a reduced parameter set, distribution $P(x)$ from Definition 4 can be written in a simpler fashion. 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 to c by R_i. Then, the following holds:

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

Example 2. Chickering, Heckerman, and Meek [5] proposed Bayesian networks "with local structure". They used a decision tree T_i (or, alternatively, a decision graph G_i) over the parent-variables of x_i for every $i \in \{1,\ldots,n\}$. 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 $R_i(\alpha) = c \in \{1,\ldots,d_i\}$ if α is routed to leaf c in T_i (or to sink c in G_i, respectively).

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 $P(x) \geq Q(x)$ and -1 otherwise. The obvious connection to concept classes in learning theory is made explicit in the following

Definition 5. *Let $\mathcal{N}$ be a Bayesian network with nodes $1, \ldots, n$ and $\mathcal{D}_{\mathcal{N}}$ 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 $\operatorname{sign}(\log(P(x)/Q(x)))$ for $P, Q \in \mathcal{D}_{\mathcal{N}}$. Note that this function attains value $+1$ if $P(x) \geq Q(x)$ and value -1 otherwise.*

The VC-dimension of $\mathcal{C}_{\mathcal{N}}$ is simply denoted as $\operatorname{VCdim}(\mathcal{N})$ throughout the paper.

2.3 Linear Arrangements in Inner Product Spaces

As explained in the introduction, we restrict ourselves to finite-dimensional Euclidean spaces and the standard scalar product $u^\top v = \sum_{i=1}^d u_i v_i$, where $u^\top$ denotes the transpose of u.

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*

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

The smallest d such that there exists a d-dimensional linear arrangement for $\mathcal{C}$ (possibly ∞ if there is no finite-dimensional arrangement) is denoted as $\operatorname{Edim}(\mathcal{C})$. [6]

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

It is easy to see that $\operatorname{Edim}(\mathcal{C}) \leq \min\{|\mathcal{C}|, |\mathcal{X}|\}$ for finite classes. Less trivial upper bounds are usually obtained constructively, by presenting an appropriate arrangement. As for lower bounds, the following is known:

Lemma 1. $\operatorname{Edim}(\mathcal{C}) \geq \operatorname{VCdim}(\mathcal{C})$.

Lemma 2 ([10]). *Let $f_1, \ldots, f_m \in \mathcal{C}$, $x_1, \ldots, x_n \in \mathcal{X}$, and $M \in \{-1,+1\}^{m \times n}$ be the matrix given by $M_{i,j} = f_i(x_j)$. Then,* $\operatorname{Edim}(\mathcal{C}) \geq \sqrt{mn}/\|M\|$*, where $\|M\| = \sup_{z \in \mathbb{R}^n : \|z\|_2 = 1} \|Mz\|_2$ denotes the spectral norm of M.*

Lemma 1 easily follows from a result by Cover [6] which states that $\operatorname{VCdim}(\{\operatorname{sign} \circ f | f \in \mathcal{F}\}) = d$ for every d-dimensional vector space $\mathcal{F}$ consisting of real-valued functions. Lemma 2 (proven in [10]) is highly non-trivial.

Let PARITY_n denote the concept class $\{h_a | a \in \{0,1\}^n\}$ on the Boolean domain given by $h_a(x) = (-1)^{a^\top x}$. Let $H_n \in \{-1,+1\}^{2^n \times 2^n}$ denote the matrix with entry $h_a(x)$ in row a and column x (Hadamard-matrix). From Lemma 2 and the well-known fact that $\|H_n\| = 2^{n/2}$ (which holds for any orthogonal matrix from $\{-1,+1\}^{2^n \times 2^n}$), one gets

Corollary 1 ([10]). $\operatorname{Edim}(\mathrm{PARITY}_n) \geq 2^n/\|H_n\| = 2^{n/2}$.

[6] Edim stands for Euclidean dimension.

3 Linear Arrangements for Bayesian Networks

In this section, we present concrete linear arrangements for several types of Bayesian networks, which leads to upper bounds on $\mathrm{Edim}(\mathcal{N})$. We sketch the proofs only.

For a set M, 2^M denotes its power set.

Theorem 1. *For every unconstrained Bayesian network, the following holds:*

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

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

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

On the right-hand side of (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 appropriate dimension is now obtained in the obvious fashion by introducing one coordinate per monomial. □

Corollary 2. *Let $\mathcal{N}_k$ denote the Bayesian network from Example 1. Then:*

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

Proof. Apply Theorem 1 and observe that

$$\bigcup_{i=1}^{n} 2^{P_i \cup \{i\}} = \bigcup_{i=k+1}^{n} \{J_i \cup \{i\} | J_i \subseteq \{i-1, \dots, i-k\}\} \cup \{J | J \subseteq \{1, \dots, k\}\} \ . \quad \square$$

Theorem 2. *Let $\mathcal{N}^R$ be a Bayesian network with a reduced parameter set $(p_{i,c})_{1 \leq i \leq n, 1 \leq c \leq d_i}$ in the sense of Definition 2. 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 in the form (2). We make use of the following obvious equation:

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

A linear arrangement of the appropriate dimension is now obtained in the obvious fashion 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). □

Theorem 3. *Let $\mathcal{N}_\sigma$ denote the logistic autoregressive Bayesian network from Definition 3. Then,* $\mathrm{VCdim}(\mathcal{N}_\sigma) = O(n^6)$.

The proof of this theorem is found in the full paper.

Remark 1. The linear arrangements for unconstrained Bayesian networks or for Bayesian networks with a reduced parameter set were easy to find. This is no accident: a similar remark is valid for every class of distributions (or densities) from the *exponential family* because (as pointed out in [8] for example) the corresponding Bayes-rule takes the form of a so-called *generalized linear rule* from which a linear arrangement is evident.[7] See the full paper for more details.

4 Lower Bounds on the Dimension of an Arrangement

In this section, we derive lower bounds on $\mathrm{Edim}(\mathcal{N})$ that match the upper bounds from Section 3 up to a small gap. Before we move on to the main results in Subsections 4.1 and 4.2, we briefly mention (without proof) some specific Bayesian networks where upper and lower bound match. The proofs are found in the full paper.

Theorem 4. $\mathrm{VCdim}(\mathcal{N}_0) = \mathrm{Edim}(\mathcal{N}_0) = n + 1$ *if $\mathcal{N}_0$ has $n \geq 2$ nodes and* $\mathrm{VCdim}(\mathcal{N}_0) = \mathrm{Edim}(\mathcal{N}_0) = 1$ *if $\mathcal{N}_0$ has 1 node.*

Theorem 5. *For $k \geq 0$, let $\mathcal{N}'_k$ denote the unconstrained network with $P_i = \{1, \ldots, k\}$ for $i = k+1, \ldots, n$ and $P_i = \emptyset$ for $i = 1, \ldots, k$. Then,* $\mathrm{VCdim}(\mathcal{N}'_k) = \mathrm{Edim}(\mathcal{N}'_k) = 2^k(n - k + 1)$

4.1 Lower Bounds Based on VC-Dimension Considerations

Since $\mathrm{VCdim}(\mathcal{C}) \leq \mathrm{VCdim}(\mathcal{C}')$ if $\mathcal{C} \leq \mathcal{C}'$, a lower bound on $\mathrm{VCdim}(\mathcal{C}')$ can be obtained from classes $\mathcal{C} \leq \mathcal{C}'$ whose VC-dimension is known or easy to determine. We first define concept classes that will fit this purpose.

Definition 7. *Let $\mathcal{N}$ be a 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 $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$. We define $\mathcal{C}_{\mathcal{N},\mathcal{F}}$ as the concept class over domain $\{0,1\}^n \setminus \{\bar{0}\}$ consisting of all functions of the form*

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

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

[7] The bound given in Theorem 1 is slightly stronger than the bound obtained from the general approach for members of the exponential family.

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.*

Lemma 3. $\mathrm{VCdim}(\mathcal{C}_{\mathcal{N},\mathcal{F}}) = \sum_{i=1}^n \mathrm{VCdim}(\mathcal{F}_i)$.

Proof. We prove 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 $\tau_i(a) := (a', 1, 0, \dots, 0)$, where $a' \in \{0,1\}^{i-1}$ is chosen such that its projection to the parent-variables of x_i coincides with a and the remaining components are projected to 0. Note that $\tau_i(a)$ is absorbed in item (x_i, f_i) of the decision list $L_{\mathcal{N},f}$. It is easy to see that the following holds. If, for $i = 1, \dots, n$, 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 first application of Lemma 3 concerns unconstrained networks.

Theorem 6. *Let $\mathcal{N}$ be an unconstrained Bayesian network and let $\mathcal{F}_i^*$ denote the set of all ±1-valued functions on domain $\{0,1\}^{m_i}$ and $\mathcal{F}^* = \mathcal{F}_1^* \times \cdots \times \mathcal{F}_n^*$. Then, $\mathcal{C}_{\mathcal{N},\mathcal{F}^*} \leq \mathcal{C}_{\mathcal{N}}$.*

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

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

An easy calculation now shows that

$$\log\left(\frac{p_{i,\alpha}}{q_{i,\alpha}}\right) = f_i(\alpha) 2^{i-1} n \text{ and } \left|\log \frac{1 - p_{i,\alpha}}{1 - q_{i,\alpha}}\right| < 1 \ . \tag{5}$$

Fix an arbitrary $x \in \{0,1\}^n \setminus \{\bar{0}\}$. 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, $L_{\mathcal{N},f}(x) = \mathrm{sign}(\log(P(x)/Q(x)))$ would follow immediately from

$$\mathrm{sign}\left(\log \frac{P(x)}{Q(x)}\right) = \mathrm{sign}\left(\log \frac{p_{i_*,\alpha_*}}{q_{i_*,\alpha_*}}\right) = f_{i_*}(\alpha_*) \ . \tag{6}$$

The second equation in (6) is evident from (5). As for the first equation in (6), we argue as follows. By the choice of i_*, $x_i = 0$ for every $i > i_*$. In combination with (3) and (5), we get

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

where $I = \{1, \ldots, n\} \setminus \{i_*\}$. 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 $2^{i_*-1}n - \sum_{j=1}^{i_*-1} 2^{j-1}n - (n-1) \geq 1$. This concludes the proof. $\square$

The next two results are straightforward applications of Lemma 3 combined with Theorems 6, 1, and Corollary 2.

Corollary 3. *For every unconstrained Bayesian network $\mathcal{N}$, the following holds:*

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

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

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

We now show that Lemma 3 can be applied in a similar fashion to the more general case of networks with a reduced parameter collection.

Theorem 7. *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 2. Let $\mathcal{F}_i^{R_i}$ denote the set of all ± 1-valued functions on domain $\{0,1\}^{m_i}$ that depend on $\alpha \in \{0,1\}^{m_i}$ only through $R_i(\alpha)$. In other words, $f \in \mathcal{F}_i^{R_i}$ iff 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 6. 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. Compare with (4). An appropriate choice is as follows:

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

The rest of the proof is completely analogous to the proof of Theorem 6. $\square$

From Lemma 3 and Theorems 7 and 2, we get

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 2. Then:*

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

Lemma 3 does not seem to apply to constrained networks. However, some of these networks allow for a similar reasoning as in the proof of Theorem 6. More precisely, the following holds:

Theorem 8. *Let $\mathcal{N}$ be a constrained Bayesian network. Assume there exists, for every $i \in \{1, \ldots, n\}$, a collection of pairwise different bit-patterns $\alpha_{i,1}, \ldots, \alpha_{i,d_i} \in \{0,1\}^{m_i}$ such that the constraints of $\mathcal{N}$ allow for the following independent decisions: for every pair (i, c), where i ranges from 1 to n and c from 1 to d_i, parameter $p_{i,\alpha_{i,c}}$ is set either to value $2^{-2^{i-1}n}/2$ or to value $1/2$. Then:*

$$\mathrm{VCdim}(\mathcal{N}) \geq \sum_{i=1}^{n} d_i \enspace .$$

Proof. For every pair (i, c), let $x_{i,c} \in \{0,1\}^n$ be the vector that has bit 1 in coordinate i, bit-pattern $\alpha_{i,c}$ in the coordinates corresponding to the parents of i, and zeros in the remaining coordinates (including positions $i+1, \ldots, n$). Following the train of thoughts in the proof of Theorem 6, it is easy to see that the vectors $x_{i,c}$ are shattered by $\mathcal{C}_{\mathcal{N}}$. □

Corollary 6. *Let $\mathcal{N}_\sigma$ denote the logistic autoregressive Bayesian network from Definition 3. Then:*

$$\mathrm{VCdim}(\mathcal{N}_\sigma) \geq \frac{1}{2}(n-1)n \enspace .$$

Proof. We aim at applying Theorem 8 with $d_i = i - 1$ for $i = 1, \ldots, n$. For $c = 1, \ldots, i-1$, let $\alpha_{i,c} \in \{0,1\}^{i-1}$ be the pattern with bit 1 in position c and zeros elsewhere. It follows now from Definition 3 that $p_{i,\alpha_{i,c}} = \sigma(w_{i,c})$. Since $\sigma(\mathbb{R}) =]0,1[$, the parameters $p_{i,\alpha_{i,c}}$ can independently be set to any value of our choice in $]0,1[$. Thus, Theorem 8 applies. □

4.2 Lower Bounds Based on Spectral Norm Considerations

We would like to show an exponential lower bound on $\mathrm{Edim}(\mathcal{N}_\sigma)$. However, at the time being, we get such a bound for a slight modification of this network only:

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 following constraints on the parameter collection:*

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

and

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

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

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

Theorem 9. *Let $\mathcal{N}'_\sigma$ denote the modified logistic autoregressive Bayesian network with $n+2$ nodes, where we assume (for sake of simplicity only) that n is a multiple of 4. Then, $\mathrm{PARITY}_{n/2} \le \mathcal{N}'_\sigma$ even if we restrict the "weights" in the parameter collection of $\mathcal{N}'_\sigma$ to integers of size $O(\log n)$.*

Proof. The mapping

$$\{0,1\}^{n/2} \ni x = (x_1, \ldots, x_{n/2}) \mapsto (\overbrace{1, x_1, \ldots, x_{n/2}, 1, \ldots, 1}^{\alpha}, 1) = x' \in \{0,1\}^{n+2} \quad , \tag{7}$$

embeds $\{0,1\}^{n/2}$ into $\{0,1\}^{n+2}$. Note that α, as indicated in (7), equals the bit-pattern of the parent-variables of x'_{n+1} (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 $x \in \{0,1\}^{n/2}$,

$$(-1)^{a^\top x} = \mathrm{sign}\left(\log \frac{P(x')}{Q(x')}\right) \quad . \tag{8}$$

(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 2-layer (unit weights) threshold circuit with $n/2$ threshold units at the first layer (and, of course, one output threshold unit at the second layer).

Fact 2 Each 2-layer threshold circuit C with polynomially bounded integer weights can be simulated by a 2-layer sigmoidal circuit C' with polynomially bounded integer weights, the same number of units, and the following output convention: $C(x) = 1 \implies C'(x) \ge 2/3$ and $C(x) = 0 \implies C'(x) \le 1/3$. The same remark holds when we replace "polynomially bounded" by "logarithmically bounded".

Fact 3 $\mathcal{N}'_\sigma$ contains (as a "substructure") a 2-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.

Fact 1 (even its generalization to arbitrary symmetric Boolean functions) is well known [16]. Fact 2 follows from a more general result by Maass, Schnitger, and Sontag. (See Theorem 4.3 in [22].) The third fact needs some explanation. (The following discussion should be compared with Definition 8.) 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 2-layer sigmoidal circuit as described in Fact 3. To this end, we set $w_{i,j} = 0$ if $1 \le i \le n/2$ or if $i,j \ge n/2+1$. We set $w_i = 0$ if $1 \le i \le n/2$. The parameters which have been set to zero are referred to as *redundant* parameters in what follows. Recall from (7) that $\alpha_0 = \alpha_{n/2+1} = \cdots = \alpha_n = 1$. From these settings (and from $\sigma(0) = 1/2$), we get

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

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

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

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

Since C' computes $(-1)^{a^\top x}$ (with the output convention from Fact 3), we get $P(x')/Q(x') \geq 2$ if $(-1)^{a^\top x} = 1$, and $P(x')/Q(x') \leq 1/2$ otherwise, which implies (8) and concludes the proof of the claim. $\square$

From Corollary 1 and Theorem 9, we get

Corollary 7. $\mathrm{Edim}(\mathcal{N}'_\sigma) \geq 2^{n/4}$.

We mentioned in the introduction (see the remarks about random projections) 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. It was shown in [10, 13] 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.

Open Problems 1) Determine Edim for the (unmodified) logistic autoregressive Bayesian network.
2) Determine Edim for other popular classes of distributions or densities (where, in the light of Remark 1, those from the exponential family look like a good thing to start with).

Acknowledgements Thanks to the anonymous referees for valuable comments and suggestions.

References

1. Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann. Hidden Markov support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, pages 3–10. AAAI Press, 2003.

2. Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40'th Annual Symposium on the Foundations of Computer Science*, pages 616–623, 1999.
3. Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. Limitations of learning via embeddings in euclidean half-spaces. *Journal of Machine Learning Research*, 3:441–461, 2002. An extended abstract of this paper appeared in the Proceedings of the 14th Annual Conference on Computational Learning Theory (COLT 2001).
4. Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.
5. David Maxwell Chickering, David Heckerman, and Christopher Meek. 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 Kaufman, 1997.
6. Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.
7. P. Mc Cullagh and J. A. Nelder. *Generalized Linear Models.* Chapman and Hall, 1983.
8. Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer Verlag, 1996.
9. Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis.* Wiley–Interscience. John Wiley & Sons, New York, 1973.
10. Jürgen Forster. A linear lower bound on the unbounded error communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002. An extended abstract of this paper appeared in the Proceedings of the 16th Annual Conference on Computational Complexity (CCC 2001).
11. Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt, and Hans Ulrich Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *Proceedings of the 21'st Annual Conference on the Foundations of Software Technology and Theoretical Computer Science*, pages 171–182, 2001.
12. Jürgen Forster, Niels Schmitt, Hans Ulrich Simon, and Thorsten Suttorp. Estimating the optimal margins of embeddings in euclidean half spaces. *Machine Learning*, 51(3):263–281, 2003. An extended abstract of this paper appeared in the Proceedings of the 14th Annual Conference on Computational Learning Theory (COLT 2001).
13. Jürgen Forster and Hans Ulrich Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. In *Proceedings of the 13th International Workshop on Algorithmic Learning Theory*, pages 128–138, 2002.
14. P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory (B)*, 44:355–362, 1988.
15. Brendan J. Frey. *Graphical Models for Machine Learning and Digital Communication.* MIT Press, 1998.
16. Andras Hajnal, Wolfgang Maass, Pavel Pudlák, Mario Szegedy, and Györgi Turán. Threshold circuits of bounded depth. *Journal of Computer and System Sciences*, 46:129–1154, 1993.
17. Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, 1998.

18. Tommi S. Jaakkola and David Haussler. Probabilistic kernel regression models. In *Proceedings of the 7th International Workshop on AI and Statistics*. Morgan Kaufman, 1999.
19. W. B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert spaces. *Contemp. Math.*, 26:189–206, 1984.
20. Eike Kiltz. On the representation of boolean predicates of the Diffie-Hellman function. In *Proceedings of 20th International Symposium on Theoretical Aspects of Computer Science*, pages 223–233, 2003.
21. Eike Kiltz and Hans Ulrich Simon. Complexity theoretic aspects of some cryptographic functions. In *Proceedings of the 9th International Conference on Computing and Combinatorics*, pages 294–303, 2003.
22. Wolfgang Maass, Georg Schnitger, and Eduardo D. Sontag. A comparison of the computational power of sigmoid and Boolean theshold circuits. In Vwani Roychowdhury, Kai-Yeung Siu, and Alon Orlitsky, editors, *Theoretical Advances in Neural Computation and Learning*, pages 127–151. Kluwer Academic Publishers, 1994.
23. Radford M. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.
24. Nuria Oliver, Bernhard Schölkopf, and Alexander J. Smola. Natural regularization from generative models. In Alexander J. Smola, Peter L. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 51–60. MIT Press, 2000.
25. Judea Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence*, pages 133–136. AAAI Press, 1982.
26. Laurence K. Saul, Tommi Jaakkola, and Michael I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.
27. Craig Saunders, John Shawe-Taylor, and Alexei Vinokourov. String kernels, Fisher kernels and finite state automata. In *Advances in Neural Information Processing Systems 15*. MIT Press, 2002.
28. Michael Schmitt. On the complexity of computing and learning with multiplicative neural networks. *Neural Computation*, 14:241–301, 2002.
29. D. J. Spiegelhalter and R. P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision support systems. *Journal of the Royal Statistical Society*, pages 35–77, 1984.
30. Koji Tsuda, Shotaro Akaho, Motoaki Kawanabe, and Klaus-Robert Müller. Asymptotic properties of the Fisher kernel. *Neural Computation*, 2003. To appear.
31. Koji Tsuda and Motoaki Kawanabe. The leave-one-out kernel. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 727–732. Springer, 2002.
32. Koji Tsuda, Motoaki Kawanabe, Gunnar Rätsch, Sören Sonnenburg, and Klaus-Robert Müller. A new discriminative kernel from probabilistic models. *Neural Computation*, 14(10):2397–2414, 2002.
33. Vladimir Vapnik. *Statistical Learning Theory*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, 1998.