Exchangeable Variable Models

Mathias Niepert   MNIEPERT@CS.WASHINGTON.EDU
Pedro Domingos   PEDROD@CS.WASHINGTON.EDU
Department of Computer Science & Engineering, University of Washington, Seattle, WA 98195, USA

Abstract

A sequence of random variables is exchangeable
if its joint distribution is invariant under variable
permutations. We introduce exchangeable variable models (EVMs) as a novel class of probabilistic models whose basic building blocks are
partially exchangeable sequences, a generalization of exchangeable sequences. We prove that a
family of tractable EVMs is optimal under zero-one loss for a large class of functions, including parity and threshold functions, and strictly
subsumes existing tractable independence-based
model families. Extensive experiments show that
EVMs outperform state of the art classifiers such
as SVMs and probabilistic models which are
solely based on independence assumptions.

1. Introduction

Conditional independence is a crucial notion that facilitates efficient inference and parameter learning in probabilistic models. Its logical and algorithmic properties as
well as its graphical representations have led to the advent
of graphical models as a discipline within artificial intelligence (Koller & Friedman, 2009). The notion of finite
(partial) exchangeability (Diaconis & Freedman, 1980a),
on the other hand, has not yet been explored as a basic building block for tractable probabilistic models. A
sequence of random variables is exchangeable if its distribution is invariant under variable permutations. Similar to conditional independence, partial exchangeability, a
generalization of exchangeability, can reduce the complexity of parameter learning and is a concept that facilitates
high tree-width graphical models with tractable inference.
For instance, the graphical models (a)-(c) with Bernoulli
variables in Figure 1 depict typical low tree-width models
based on the notion of (conditional) independence. Graphical models (d)-(f) have high tree-width but are tractable

if we assume the variables with identical shades to be exchangeable. We will see that EVMs are especially beneficial for high-dimensional and sparse domains such as
text and collaborative filtering problems. While there exists
work on tractable models, with a majority focusing on low
tree-width graphical models, a framework for finite partial
exchangeability as a basic building block of *tractable* probabilistic models seems natural but does not yet exist.
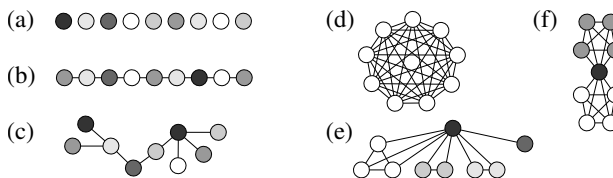

Figure 1. Illustration of low tree-width models exploiting independence (a)-(c) and exchangeable variable models (EVMs)
exploiting finite exchangeability (variable nodes with identical
shades are exchangeable) (d)-(f).

We propose exchangeable variable models (EVMs), a
novel family of probabilistic models for classification and
probability estimation. While most probabilistic models
are built on the notion of conditional independence and
its graphical representation, EVMs have finite partially exchangeable sequences as basic components. We show that
EVMs can represent complex positive and negative correlations between large sets of variables with few parameters
and without sacrificing tractable inference. The parameters of EVMs are estimated under the maximum-likelihood
principle and we assume the examples to be independent
and identically distributed. We develop methods for efficient probabilistic inference, maximum-likelihood estimation, and structure learning.

We introduce the mixtures of EVMs (MEVMs) family of
models which is strictly more expressive than the naive
Bayes family of models but as efficient to learn. MEVMs
represent classifiers that are optimal under zero-one loss
for a large class of Boolean functions including parity and
threshold functions. Extensive experiments show that exchangeable variable models, when combined with the notion of conditional independence, are effective both for

classification and probability estimation. The MEVM classifier significantly outperforms state of the art classifiers on numerous high-dimensional and sparse data sets. MEVMs also outperform several tractable graphical model classes on typical probability estimation problems while being orders of magnitudes more efficient.

2. Background

We begin by reviewing the statistical concepts of finite exchangeability and finite partial exchangeability.

2.1. Finite Exchangeability

Finite exchangeability is best understood in the context of a finite sequence of binary random variables such as a finite number of coin tosses. Here, finite exchangeability means that it is only the number of heads that matters and not their particular order. Since exchangeable variables are not necessarily independent, finite exchangeability can model highly correlated variables, a graphical representation of which would be the fully connected graph with high treewidth (see Figure 1(d)). However, as we will later see, the number of parameters and the complexity of inference remains linear in the number of variables.

Definition 2.1 (Exchangeability). Let $X_1, ..., X_n$ be a sequence of random variables with joint distribution P and let $S(n)$ be the group of all permutations acting on $\{1, ..., n\}$. We say that $X_1, ..., X_n$ is exchangeable if $P(X_1, ..., X_n) = P(X_{\pi(1)}, ..., X_{\pi(n)})$ for all $\pi \in S(n)$.

In this paper, we are concerned with exchangeable *variables* and iid *examples*. The literature has mostly focused on exchangeability of an *infinite* sequence of random variables. In this case, one can express the joint distribution as a mixture of iid sequences (de Finetti, 1938). However, for finite sequences of exchangeable variables this representation is inadequate – while finite exchangeable sequences can be approximated with de Finetti style mixtures of iid sequences, these approximations are not suitable for finite sequences of random variables not extendable to an infinite exchangeable sequence (Diaconis & Freedman, 1980b). Moreover, negative correlations can only be modeled in the finite case. There are interesting connections between the automorphisms of graphical models and finite exchangeability (Niepert, 2012). An alternative approach to exchangeability considers its relationship to sufficiency (Diaconis & Freedman, 1980a; Lauritzen et al., 1984) which is at the core of our work.

2.2. Finite Partial Exchangeability

The assumption that all variables of a probabilistic model are exchangeable is often too strong. Fortunately, finite exchangeability can be generalized to the concept of finite
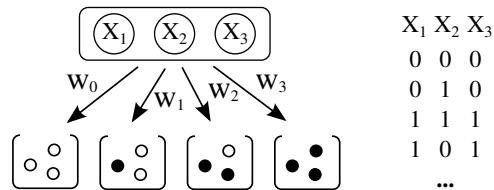

Figure 2. A finite sequence of exchangeable variables can be parameterized as a unique mixture of urn processes. Each such urn process is a series of draws without replacement.

partial exchangeability using the notion of a statistic.

Definition 2.2 (Partial Exchangeability). Let $X_1, ..., X_n$ be a sequence of random variables with distribution P, let $\mathbf{Val}(X_i)$ be the domain of X_i, and let $\mathcal{T}$ be a finite set. The sequence $X_1, ..., X_n$ is partially exchangeable with respect to the statistic $T : \mathbf{Val}(X_1) \times ... \times \mathbf{Val}(X_n) \to \mathcal{T}$ if

$$T(\mathbf{x}) = T(\mathbf{x}') \text{ implies } P(\mathbf{x}) = P(\mathbf{x}'),$$

where $\mathbf{x}$ and $\mathbf{x}'$ are assignments to the sequence of random variables $X_1, ..., X_n$.

The following theorem states that the joint distribution of a sequence of random variables, which is partially exchangeable with respect to a statistic T, is a unique mixture of uniform distributions.

Theorem 2.3. (Diaconis & Freedman, 1980a) *Let $X_1, ..., X_n$ be a sequence of random variables with distribution P, let $\mathcal{T}$ be a finite set, and let $T : \mathbf{Val}(X_1) \times ... \times \mathbf{Val}(X_n) \to \mathcal{T}$ be a statistic. Moreover, let $S_t = \{\mathbf{x} \in \mathbf{Val}(X_1) \times ... \times \mathbf{Val}(X_n) \mid T(\mathbf{x}) = t\}$, let U_t be the uniform distribution on S_t, and let $w_t = P(S_t)$. If $X_1, ..., X_n$ is partially exchangeable with respect to T, then*

$$P(\mathbf{x}) = \sum_{t \in \mathcal{T}} w_t U_t(\mathbf{x}). \tag{1}$$

The theorem provides an implicit description of the distributions U_t. The challenge for specific families of random variables lies in finding a statistic T with respect to which a sequence of variables is partially exchangeable and an efficient algorithm to compute the probabilities $U_t(\mathbf{x})$. For the case of exchangeable sequences of discrete random variables and, in particular, exchangeable sequences of binary random variables, an explicit description does exist and is well-known in the statistics literature (Diaconis & Freedman, 1980a; Stefanescu & Turnbull, 2003).

Example 2.4. Let X_1, X_2, X_3 be three exchangeable binary variables with joint distribution P. Then, the sequence X_1, X_2, X_3 is partially exchangeable with respect to the statistic $T : \{0, 1\}^3 \to \mathcal{T} = \{0, 1, 2, 3\}$ with $T(\mathbf{x} = (x_1, x_2, x_3)) = x_1 + x_2 + x_3$. Thus, we can write

$$P(\mathbf{x}) = \sum_{t \in \mathcal{T}} w_t U_t(\mathbf{x}),$$

where $w_t = P(T(\mathbf{x}) = t)$, $U_t(\mathbf{x}) = [[T(\mathbf{x}) = t]]\binom{3}{t}^{-1}$, and $[[\cdot]]$ is the indicator function. Hence, the distribution can be parameterized as a unique mixture of four urn processes, where T's value is the number of black balls. Figure 2 illustrates the mixture model. The generative process is as follows. First, choose one of the four urns according to the mixing weights w_t; then draw three consecutive balls from the chosen urn without replacement.

3. Exchangeable Variable Models

We propose exchangeable variable models (EVMs) as a novel family of tractable probabilistic models for classification and probability estimation. While probabilistic graphical models are built on the notion of (conditional) independence and its graphical representation, EVMs are built on the notion of finite (partial) exchangeability. EVMs can model both *negative* and *positive* correlations in what would be high tree-width graphical models without losing tractability of probabilistic inference.

The basic components of EVMs are tuples $(\mathbf{X}, T)$ where $\mathbf{X}$ is a sequence of *discrete* random variables partially exchangeable with respect to the statistic T with values $\mathcal{T}$.

3.1. Probabilistic Inference

We can relate finite partial exchangeability to tractable probabilistic inference (see also (Niepert & Van den Broeck, 2014)). We assume that for every joint assignment $\mathbf{x}$, $P(\mathbf{x})$ can be computed in time $\mathbf{poly}(|\mathbf{X}|)$.

Proposition 3.1. *Let $\mathbf{X}$ be partially exchangeable with respect to the statistic T with values $\mathcal{T}$, let $|\mathcal{T}| = \mathbf{poly}(|\mathbf{X}|)$, and let, for any partial assignment $\mathbf{e}$, $S_{t,\mathbf{e}} := \{\mathbf{x} \mid T(\mathbf{x}) = t \text{ and } \mathbf{x} \sim \mathbf{e}\}$, where $\mathbf{x} \sim \mathbf{e}$ denotes that $\mathbf{x}$ and $\mathbf{e}$ agree on the variables in their intersection (Koller & Friedman, 2009). If we can in time $\mathbf{poly}(|\mathbf{X}|)$,*

(1) for every $\mathbf{e}$ and every $t \in \mathcal{T}$, decide if there exists an $\mathbf{x} \in S_{t,\mathbf{e}}$ and, if so, construct such an $\mathbf{x}$,

then the complexity of MAP inference, that is, computing $\arg\max_{\mathbf{y}} P(\mathbf{y}, \mathbf{e})$ for any partial assignment $\mathbf{e}$, is $\mathbf{poly}(|\mathbf{X}|)$. If, in addition, we can in time $\mathbf{poly}(|\mathbf{X}|)$,

(2) for every $\mathbf{e}$ and every $t \in \mathcal{T}$, compute $|S_{t,\mathbf{e}}|$,

then the complexity of marginal inference, that is, computing $P(\mathbf{e})$ for any partial assignment $\mathbf{e}$, is $\mathbf{poly}(|\mathbf{X}|)$.

Proposition 3.1 generalizes to probabilistic models where $P(\mathbf{x})$ can only be computed up to a constant factor Z such as undirected graphical models. Please note that computing conditional probabilities is tractable whenever conditions (1) and (2) are satisfied. We say a statistic is tractable if either of the conditions is fulfilled.

Proposition 3.1 provides a theoretical framework for developing tractable non-local potentials. For instance, for n exchangeable Bernoulli variables, the complexity of MAP and marginal inference is polynomial in n. This follows from the statistic T satisfying conditions (1) and (2) and since $|\mathcal{T}| = n + 1$. Related work on cardinality-based potentials has mostly focused on MAP inference (Gupta et al., 2007; Tarlow et al., 2010). Finite exchangeability also speaks to marginal inference via the tractability of computing $U_t(\mathbf{e}) = |S_{t,\mathbf{e}}|^{-1}$. EVMs can model unary potentials using singleton sets of exchangeable variables. While not all instances of finite partial exchangeability result in tractable probabilistic models there exist several examples satisfying conditions (1) and (2) which go beyond finite exchangeability. In the supplementary material, in addition to the proofs of all theorems and propositions, we present examples of tractable statistics that are different from those associated with cardinality-based potentials (Gupta et al., 2007; Tarlow et al., 2010; 2012; Bui et al., 2012).

3.2. Parameter Learning

The parameters of finite sequences of partially exchangeable variables are the mixture weights of the parameterization given in Equation 1 of Theorem 2.3. Estimating the parameters of these basic components of EVMs is a crucial task. We derive the maximum-likelihood estimates for these mixture weight vectors.

Theorem 3.2. *Let $X_1, ..., X_n$ be a sequence of random variables with joint distribution P, let T be a statistic with distinct values $t_0, ..., t_k$, and let $X_1, ..., X_n$ be partially exchangeable with respect to T. The ML estimates for N examples, $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$, are $\mathtt{MLE}[(w_0, ..., w_k)] = \left(\frac{c_0}{N}, ..., \frac{c_k}{N}\right)$, where $c_i = \sum_{j=1}^{N}[[T\left(\mathbf{x}^{(j)}\right) = t_i]]$.*

Hence, the statistical parameters to be estimated are identical to the statistical parameters of a multinomial distribution with $|\mathcal{T}|$ distinct categories.

3.3. Structure Learning

Let $\hat{\mathbf{X}}$ be a sequence of random variables and let $\hat{\mathbf{x}}^{(1)}, ..., \hat{\mathbf{x}}^{(N)}$ be N iid examples drawn from the data-generating distribution. In order to learn the structure of EVMs we need to address two problems.

Problem 1: Find subsequences $\mathbf{X} \subseteq \hat{\mathbf{X}}$ that are exchangeable with respect to a given tractable statistic T. This identifies individual EVM components $(\mathbf{X}, T)$ for which tractable inference and learning is possible. We may utilize different tractable statistics for different components.

Problem 2: Construct graphical models whose potentials are the previously learned tractable EVM components. In order to preserve tractability of the global model, we have to restrict the class of possible graphical structures.

We now present approaches to these two problems that learn expressive EVMs while maintaining tractability.

Let us first address **Problem 1**. We focus on EVMs with finitely exchangeable components. Fortunately, there exist several *necessary* conditions for finite exchangeability (see Definition 2.1) of a sequence of random variables.

Proposition 3.3. *The following statements are necessary conditions for exchangeability of a finite sequence of random variables* $X_1, ..., X_n$. *For all* $i, j, i', j' \in \{1, ..., n\}$ *with* $i \neq j$ *and* $i' \neq j'$

(1) $\mathbf{E}(X_i) = \mathbf{E}(X_j)$;

(2) $\mathbf{Var}(X_i) = \mathbf{Var}(X_j)$; *and*

(3) $\mathbf{Cov}(X_i, X_j) = \mathbf{Cov}(X_{i'}, X_{j'}) \geq -\frac{\mathbf{Var}(X_i)}{(n-1)}$.

The necessary conditions can be exploited to assess whether a sequence of variables is finitely exchangeable. In order to learn EVM components $(\mathbf{X}, T)$ we assume that a sequence of variables is exchangeable unless a statistical test contradicts some or all of the necessary conditions for finite exchangeability. For instance, if a statistical test deemed the expectations $\mathbf{E}(X)$ and $\mathbf{E}(X')$ for two variables X and X' identical, we could assume X and X' to be exchangeable. If we wanted the statistical test for finite exchangeability to be more specific and less sensitive, we would also require conditions (2) and/or (3) to hold. Please note the analogy to structure learning with conditional independence tests. Instead of identifying (conditional) independencies we identify finite exchangeability among random variables. For a sequence of identically distributed variables the assumption of exchangeability is *weaker* than that of independence. Testing whether two discrete variables have identical mean and variance is efficient algorithmically. Of course, the application of the necessary conditions for finite exchangeability is only one possible approach to learning the components of EVMs.

Let us now turn to **Problem 2**. To ensure tractability, the global graphical structure has to be restricted to tractable classes such as chains and trees. Here, we focus on mixture models where, conditioned on the values of the latent variable, $\hat{\mathbf{X}}$ is partitioned into exchangeable blocks (see Figure 3). Hence, for each value y of the latent variable, we perform the statistical tests of **Problem 1** with estimates of the conditional expectations $\mathbf{E}(X \mid y)$. We introduce this class of EVMs in the next section and leave more complex structures to future work.

In the context of longitudinal studies and repeated-measures experiments, where an observation is made at different times and under different conditions, there exist several models taking into account the correlation between these observations and assuming identical or similar
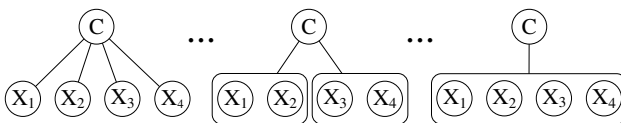

Figure 3. The combination of exchangeable and independent variables leads to a spectrum of models. On the one end is the model where, conditioned on the class, all variables are independent (but possibly not identically distributed; left). On the other end is the model where, conditioned on the class, all variables are exchangeable (but possibly correlated; right). The partition of the variables into exchangeable blocks can vary with the class value.

covariance structure for subsets of the variables (Jennrich & Schluchter, 1986). These compound symmetry models, however, do not make the assumption of exchangeability and, therefore, do not generally facilitate tractable inference. Nevertheless, finite exchangeability can be seen as a form of parameter tying, a method that has also been applied in the context of hidden Markov models, neural networks (Rumelhart et al., 1986) and, most notably, statistical relational learning (Getoor & Taskar, 2007). Collective graphical models (Sheldon & Dietterich, 2011) (CGMs) and high-order potentials (Tarlow et al., 2010; 2012) (HOPs) are models based on non-local potentials. Proposition 3.3 can be applied for learning the structure of novel tractable instances of CGMs and HOPs.

4. Exchangeable Variable Models for Classification and Probability Estimation

We are now in the position to design model families that combine the notions of (partial) exchangeability with that of (conditional) independence. Instead of specifying a structure that solely models the (conditional) independence characteristics of the probabilistic model, EVMs also specify sequences of variables that are (partially) exchangeable. The previous results provide the necessary tools to learn both the structure and parameters of partially exchangeable sequences and to perform tractable probabilistic inference.

The possibilities for building families of exchangeable variable models (EVMs) are vast. Here, we focus on a family of mixtures of EVMs generalizing the widely used naive Bayes model. The family of probabilistic models is therefore also related to research on extending the naive Bayes classifier (Domingos & Pazzani, 1997; Rennie et al., 2003). The motivation behind this novel class of EVMs is that it facilitates *both* tractable maximum-likelihood learning *and* tractable probabilistic inference.

In line with existing work on mixture models, we derive the maximum-likelihood estimates for the fully observed setting, that is, when there are no examples with missing class labels. We also discuss the expectation maximization (EM)

algorithm for the case where the data is partially observed, that is, when examples with missing class labels exist.

Definition 4.1 (Mixture of EVMs). The mixture of EVMs (MEVM) model consists of a class variable Y with k possible values, a set of binary attributes $\hat{\mathbf{X}} = \{X_1, ..., X_n\}$ and, for each $y \in \{1, ..., k\}$, a set $\mathcal{X}_y$ specifying a partition of the attributes into blocks of exchangeable sequences. The structure of the model, therefore, is defined by $\mathcal{X} = \{\mathcal{X}_i\}_{i=1}^{k}$, the set of attribute partitions, one for each class value. The model has the following parameters:

1. A parameter $p(y)$ for every $y \in \{1, ..., k\}$ specifying the prior probability of seeing class value y.

2. A parameter $q_{(\mathbf{X})}(\ell \mid y)$ for every $y \in \{1, ..., k\}$, every $\mathbf{X} \in \mathcal{X}_y$, and every $\ell \in \{0, 1, ..., |\mathbf{X}|\}$. The value of $q_{(\mathbf{X})}(\ell \mid y)$ is the probability of the exchangeable sequence $\mathbf{X} \subseteq \hat{\mathbf{X}}$ having an assignment with ℓ number of 1s, conditioned on the class label being y.

Let $\mathtt{n}_{\mathbf{X}}(\hat{\mathbf{x}})$ be the number of 1s in the joint assignment $\hat{\mathbf{x}}$ projected onto the variable sequence $\mathbf{X} \subseteq \hat{\mathbf{X}}$. The probability for every $y, \hat{\mathbf{x}} = (x_1, ..., x_n)$ is then defined as

$$\mathbf{P}(y, \hat{\mathbf{x}}) = p(y) \prod_{\mathbf{X} \in \mathcal{X}_y} q_{(\mathbf{X})}(\mathtt{n}_{\mathbf{X}}(\hat{\mathbf{x}}) \mid y) \binom{|\mathbf{X}|}{\mathtt{n}_{\mathbf{X}}(\hat{\mathbf{x}})}^{-1}.$$

Hence, conditioned on the class, the attributes are partitioned into mutually independent and disjoint blocks of exchangeable sequences. Figure 3 illustrates the model family with the naive Bayes model being positioned on one end of the spectrum. Here, $\mathcal{X}_y = \{\{X_1\}, ..., \{X_n\}\}$ for all $y \in \{1, ..., k\}$. On the other end of the spectrum is the model that assumes full exchangeability conditioned on the class. Here, $\mathcal{X}_y = \{\{X_1, ..., X_n\}\}$ for all $y \in \{1, ..., k\}$. For binary attributes, the number of *free* parameters is $k + kn - 1$ for *each* member of the MEVM family. The following theorem provides the maximum-likelihood estimates for these parameters.

Theorem 4.2. *The maximum-likelihood estimates for a MEVM with attributes $\hat{\mathbf{X}}$, structure $\mathcal{X} = \{\mathcal{X}_i\}_{i=1}^{k}$, and a sequence of examples $\left(y^{(i)}, \hat{\mathbf{x}}^{(i)}\right), 1 \leq i \leq N$, are*

$$p(y) = \frac{\sum_{i=1}^{N} [[y^{(i)} = y]]}{N}$$

and, for each y and each $\mathbf{X} \in \mathcal{X}_y$,

$$q_{(\mathbf{X})}(\ell \mid y) = \frac{\sum_{i=1}^{N} [[y^{(i)} = y \text{ and } \mathtt{n}_{\mathbf{X}}(\hat{\mathbf{x}}^{(i)}) = \ell]]}{\sum_{i=1}^{N} [[y^{(i)} = y]]}.$$

We utilize MEVMs for classification problems by learning the parameters and computing the MAP state of the

---

Algorithm 1 Expectation Maximization for MEVMs

Input: The number of classes k. Training examples $\langle \hat{\mathbf{x}}^{(i)} = (x_1^{(i)}, ..., x_n^{(i)}) \rangle, 1 \leq i \leq N$. A parameter specifying a stopping criterion.

Initialization: Assign $\lfloor N/k \rfloor$ random examples to each mixture component. For each class value $y \in \{1, ..., k\}$, partition the n variables into exchangeable sequences $\mathcal{X}_y^{(0)}$, and compute $p^{(0)}(y)$ and $q_{(\mathbf{X})}^{(0)}(\ell \mid y)$ for each $\mathbf{X} \in \mathcal{X}_y^{(0)}$ and $0 \leq \ell \leq |\mathbf{X}|$ using Theorem 4.2.

Iterate: until stopping criterion is met

1. For $i = 1, ..., N$ and $y = 1, ..., k$ compute

$$\delta(y \mid i) = \frac{\mathbf{P}^{(t-1)}(y, \hat{\mathbf{x}}^{(i)})}{\sum_{j=1}^{k} \mathbf{P}^{(t-1)}(j, \hat{\mathbf{x}}^{(i)})}.$$

2. For each $y \in \{1, ..., k\}$, partition the variables into blocks of exchangeable sequences $\mathcal{X}_y^{(t)}$.

3. Update parameters for both $\mathcal{X}_y^{(t-1)}$ and $\mathcal{X}_y^{(t)}$:

$$p^{(t)}(y) = \frac{\sum_{i=1}^{N} \delta(y \mid i)}{N},$$

$$q_{(\mathbf{X})}^{(t)}(\ell \mid y) = \frac{\sum_{i=1}^{N} [[\mathtt{n}_{\mathbf{X}}(\hat{\mathbf{x}}^{(i)}) = \ell]] \, \delta(y \mid i)}{\sum_{i=1}^{N} \delta(y \mid i)}.$$

4. Select the new block structure according to the maximum log-likelihood on training examples.

Output: Structure and parameter estimates.

---

class variable conditioned on assignments to the attribute variables. For probability estimation the class is latent and we can apply Algorithm 1. The expectation maximization (EM) algorithm is initialized by assigning random examples to the mixture components. In each EM iteration, the examples are fractionally assigned to the components, and the block structure and parameters are updated. Finally, either the previous or current structure is chosen based on the maximum likelihood. For the structure learning step we can, for instance, apply conditions from Proposition 3.3 where we use the conditional expectations $\mathbf{E}(X_j \mid y)$, estimated by $\sum_{i=1}^{N} x_j^{(i)} \delta(y \mid i)/N$, for the statistical tests to construct $\mathcal{X}_y$. Since the new structure is chosen from a set containing the structure from the previous EM iteration, the convergence of Algorithm 1 follows from that of structural expectation maximization (Friedman, 1998).

A crucial question is how *expressive* the novel model family is. We provide an analytic answer by showing that MEVMs are globally optimal under zero-one loss for a large class of Boolean functions, namely, conjunctions and disjunctions of attributes and symmetric Boolean functions. *Symmetric Boolean functions* are Boolean function whose value depends only on the number of ones in the

input (Canteaut & Videau, 2005). The class includes (a) Threshold functions, whose value is 1 on inputs vectors with k or more ones for a fixed k; (b) Exact-value functions, whose value is 1 on inputs vectors with k ones for a fixed k; (c) Counting functions, whose value is 1 on inputs vectors with the number of ones congruent to $k \mod m$ for fixed k, m; and (d) Parity functions, whose value is 1 if the input vector has an odd number of ones.

Definition 4.3. (Domingos & Pazzani, 1997) The Bayes rate for an example is the lowest zero-one loss achievable by any classifier on that example. A classifier is *locally optimal* for an example iff its zero-one loss on that example is equal to the Bayes rate. A classifier is *globally optimal* for a sample iff it is locally optimal for every example in that sample. A classifier is globally optimal for a problem iff it is globally optimal for all possible samples of that problem.

We can now state the following theorem.

Theorem 4.4. *The mixtures of EVMs family is globally optimal under zero-one loss for*

1. *Conjunctions and disjunctions of attributes;*

2. *Symmetric Boolean functions such as*

   - *Threshold (m-of-n) functions*
   - *Parity functions*
   - *Counting functions*
   - *Exact value functions*

Theorem 4.4 is striking as the parity function and its special case, the XOR function, are instances of not linearly separable functions which are often used as examples of particularly challenging classification problems. The optimality for symmetric Boolean functions holds even for the model that assumes *full* exchangeability of the attributes given the value of the class variable (see Figure 3, right). It is known that the naive Bayes classifier is *not* globally optimal for threshold (m-of-n) functions despite them being linearly separable (Domingos & Pazzani, 1997). Hence, combining conditional independence and exchangeability leads to highly tractable probabilistic models that are globally optimal for a broader class of Boolean functions.

5. Experiments

We conducted extensive experiments to assess the efficiency and effectiveness of MEVMs as tractable probabilistic models for classification and probability estimation. A major objective is the comparison of MEVMs and naive Bayes models. We also compare MEVMs with several state of the art classification algorithms. For the probability estimation experiments, we compare MEVMs to latent naive Bayes models and several widely used tractable graphical model classes such as latent tree models.

Table 1. Properties of the classification data sets and mean and standard deviation of the number of MEVM blocks.

Data set	$|V|$	Train	Test	Blocks
Parity	1,000	10^6	10,000	1.3 ± 0.3
Counting	1,000	10^6	10,000	1.9 ± 0.9
M-of-n	1,000	10^6	10,000	2.4 ± 1.6
Exact	1,000	10^6	10,000	3.2 ± 2.1
20Newsgrp	19,726.1	1,131.4	753.2	19.2 ± 1.5
Reuters-8	19,398.0	1,371.3	547.2	16.9 ± 9.1
Polarity	38,045.8	1,800.0	200.0	34.1 ± 0.7
Enron	43,813.6	4,000.0	1,000.0	30.2 ± 6.0
WebKB	7,290.0	1,401.5	698.0	19.3 ± 3.6
MNIST	784.0	12,000.0	2,000.0	72.3 ± 3.1

5.1. Classification

We evaluated the MEVM classifier using both synthetic and real-world data sets. Each synthetic data set consists of 10^6 training and 10000 test examples. Let $\mathbf{n}(\mathbf{x})$ be the number of ones of the example $\mathbf{x}$. The parity data was generated by sampling uniformly at random an example $\mathbf{x}$ from the set $\{0,1\}^{1000}$ and assigning it to the first class if $\mathbf{n}(\mathbf{x}) \mod 2 = 1$, and to the second class otherwise. For the 10-of-1000 data set we assigned an example $\mathbf{x}$ to the first class if $\mathbf{n}(\mathbf{x}) \geq 10$, and to the second class otherwise. For the counting data set we assigned an examples $\mathbf{x}$ to the first class if $\mathbf{n}(\mathbf{x}) \mod 5 = 3$, and to the second class otherwise. For the exact data set we assigned an example $\mathbf{x}$ to the first class if $\mathbf{n}(\mathbf{x}) \in \{0, 200, 400, 600, 800, 1000\}$, and to the second class otherwise.

We used the SCIKIT 0.14^1 functions to load the 20Newsgroup train and test samples. We removed headers, footers, and quotes from the training and test documents. This renders the classification problem more difficult and leads to significantly higher zero-one loss for all classifiers. For the Reuters-8 data set we considered only the Reuters-21578 documents with a single topic and the top 8 classes that have at least one train and one test example. For the WebKB text data set we considered the classes `project`, `course`, `faculty`, and `student`. For all text data sets we used the binary bag-of-word representation resulting in feature spaces with up to 45000 dimensions. For the MNIST data set, a collection of hand-written digits, we set a feature value to 1 if the original feature value was greater than 50, and to 0 otherwise. The polarity data set is a well-known sentiment analysis problem based on movie reviews (Pang & Lee, 2004). The problem is to classify movie reviews as either positive or negative. We used the cross-validation splits provided by the authors. The Enron spam data set is a collection of e-mails from the Enron corpus that was divided into spam and no-spam messages (Metsis et al.,

---

[1] http://scikit-learn.org/

Table 2. Accuracy values for the two-class experiments. Bold numbers indicate significance (paired t-test; $p < 0.01$) compared to non-bold results in the same row.

Data set	MEVM	NB	DT	SVM	5-NN
Parity	**0.958**	0.497	0.501	0.493	0.502
Counting	**0.967**	0.580	0.655	0.768	0.765
M-of-n	**0.994**	0.852	0.990	**0.995**	0.715
Exact	**0.996**	0.566	0.983	**0.995**	0.974
20Newsgrp	**0.905**	0.829	0.803	0.867	0.582
Reuters-8	**0.968**	0.940	**0.965**	**0.982**	0.881
Polarity	0.826	0.794	0.623	**0.859**	0.520
Enron	**0.980**	0.915	0.948	0.972	0.743
WebKB	**0.943**	0.907	0.899	**0.952**	0.780
MNIST	0.969	0.964	0.981	0.983	**0.995**

Table 3. Accuracy values for the multi-class experiments. Bold numbers indicate significance (paired t-test; $p < 0.01$) compared to non-bold results in the same column.

Classifier	20Newsgrp	Reuters-8	WebKB	MNIST
MEVM	**0.626**	**0.911**	**0.860**	**0.855**
NB	0.537	0.862	0.783	0.842

2006). Here, we applied randomized 100-fold validation. We did not apply feature extraction algorithms to any of the data sets. Table 1 lists the properties of the data sets and the mean and standard deviation of the number of blocks of the MEVMs. We distinguished between two-class and multi-class (more than 2 classes) problems. When the original data set had more than two classes, we created the two-class problems by considering every pair of classes as a separate cross-validation problem. We draw this distinction because we want to compare classification approaches independent of particular multi-class strategies (1-vs-n, 1-vs-1, etc.).

We exploited necessary condition (1) from Proposition 3.3 to learn the block structure of the MEVM classifiers. For each pair of variables X, X' and each class value y, we applied Welch's t-test to test the null hypothesis $\mathbf{E}(X \mid y) = \mathbf{E}(X' \mid y)$. If, for two variables, the test's p-value was less than 0.1, we rejected the null hypothesis and placed them in different blocks conditioned on y. We applied Laplace smoothing with a constant of 0.1. The same parameter values were applied across *all* data sets and experiments. For all other classifiers we used the SCIKIT 0.14 implementations naive_bayes.BernoulliNB, tree.DecisionTreeClassifier, svm.LinearSVC, and neighbors.KNeighborsClassifier. We used the classifiers' standard settings except for the naive Bayes classifier where we applied a Laplace smoothing constant (alpha) of 0.1 to ensure a fair comparison (NB results deteriorated for alpha values of 1.0 and 0.01). The standard setting for the classifiers are available as part of the SCIKIT 0.14 documentation. All implementations and data sets will be published.

Table 2 lists the results for the two-class problems. The MEVM classifier was one of the best classifiers for 8 out of the 10 data sets. With the exception of the MNIST data set, where the difference was insignificant, MEVM significantly outperformed the naive Bayes classifier (NB) on all data sets. The MEVM classifier outperformed SVMs on 4 data sets, two of which are real-world text classification problems and achieved a tie on 4. For the parity data set only the MEVM classifier was better than random. Table 3 shows the results on the multi-class problems. Here, the MEVM classifier significantly outperforms naive Bayes on all data set. The MEVM classifier outperformed all classifiers on the 20Newsgroup and was a close second on the Reuters-8 and WebKB data sets. The MEVM classifier is particularly suitable for high-dimensional and sparse data sets. We hypothesize that this has three reasons. First, MEVMs can model both negative and positive correlations between variables. Second, MEVMs perform a non-linear transformation of the feature space. Third, MEVMs cluster noisy variables into blocks of exchangeable sequences which acts as a form of regularization in sparse domains.

5.2. Probability Estimation

We conducted experiments with a widely used collection of data sets (Van Haaren & Davis, 2012; Gens & Domingos, 2013; Lowd & Rooshenas, 2013). Table 4 lists the number of variables, training and test examples, and the number of blocks of the MEVM models. We set the latent variable's domain size to 20 for each problem and applied the same EM initialization for MEVMs and NB models. This way we could compare NB and MEVM independent of the tuning parameters specific to EM. We implemented EM exactly as described in Algorithm 1. For step (2), we exploited Proposition 3.3 (1) and, for each y, partitioned the variables into exchangeable blocks by performing a series of Welch's t-tests on the expectations $\mathbf{E}(X_j \mid y)$, estimated by $\sum_{i=1}^{N} x_j^{(i)} \delta(y \mid i)/N$, assigning two variables to different blocks if the null hypothesis of identical means could be rejected at a significance level of 0.1. For MEVM and NB we again used a Laplace smoothing constant of 0.1. We ran EM until the average log-likelihood increase between iterations was less than 0.001. We restarted EM 10 times and chose the model with the maximal log-likelihood on the training examples. We did not use the validation data. For LTM (Choi et al., 2011), we applied the four methods, CLRG, CLNJ, regCLRG, and regCLNJ, and chose the model with the highest validation log-likelihood.

Table 5 lists the average log-likelihood of the test data for the MEVM, the latent naive Bayes (Lowd & Domingos, 2005) (NB), the latent tree (LTM), and the Chow-Liu tree model (Chow & Liu, 2006) (CL). Even without exploiting the validation data for model tuning, the MEVM models outperformed the CL models on all, and the LTMs on

Table 4. Properties of the data sets used for probability estimation and mean and standard deviation of the number of MEVM blocks.

Data set	$|V|$	Train	Test	Blocks
NLTCS	16	16,181	3,236	8.8 ± 1.9
MSNBC	17	291,326	58,265	15.9 ± 1.1
KDDCup 2000	64	180,092	34,955	15.8 ± 4.7
Plants	69	17,412	3,482	15.9 ± 2.9
Audio	100	15,000	3,000	13.7 ± 3.0
Jester	100	9,000	4,116	10.4 ± 2.0
Netflix	100	15,000	3,000	14.8 ± 3.2
MSWeb	294	29,441	5,000	21.3 ± 2.0
Book	500	8,700	1,739	12.4 ± 2.9
WebKB	839	2,803	838	10.6 ± 2.3
Reuters-52	889	6,532	1,540	16.7 ± 3.1
20Newsgroup	910	11,293	3,764	17.9 ± 3.7

Table 5. Average log-likelihood of the MEVM, the naive Bayes, the latent tree, and the Chow-Liu tree model.

Data set	MEVM	NB	LTM	CL
NLTCS	-6.04	-6.04	-6.46	-6.76
MSNBC	**-6.23**	-6.71	-6.52	-6.54
KDDCup 2000	**-2.13**	-2.15	-2.18	-2.29
Plants	**-14.86**	-15.10	-16.39	-16.52
Audio	**-40.63**	-40.69	-41.89	-44.37
Jester	-53.22	**-53.19**	-55.17	-58.23
Netflix	**-57.84**	-57.87	-58.53	-60.25
MSWeb	-9.96	-9.96	-10.21	-10.19
Book	-34.63	-34.80	**-34.23**	-34.70
WebKB	-157.21	-158.01	**-156.84**	-163.48
Reuters-52	**-86.98**	-87.32	-91.25	-94.37
20Newsgroup	**-152.69**	-152.78	-156.77	-164.13

all but two of the data set. MEVMs achieve the highest log-likelihood score on 7 of the 12 data sets. With the exception of the Jester data set, MEVMs either outperformed or tied the NB model. While the results indicate that MEVMs are effective for higher-dimensional and sparse data sets, where the increase in log-likelihood was most significant, MEVMs also outperformed the NB models on 3 data sets with less than 100 variables. The MEVM and NB models have exactly the same number of free parameters. Since results on the same data sets are available for other tractable model classes we also compared MEVMs with SPNs (Gens & Domingos, 2013) and ACMNs (Lowd & Rooshenas, 2013). Here, MEVMs are outperformed by the more complex SPNs on 5 and by ACMNs on 6 data sets. However, MEVMs are competitive and outperform SPNs on 7 and ACMNs on 6 of the 12 data sets. Following previous work (Van Haaren & Davis, 2012), we applied the Wilcoxon signed-rank test. MEVM outperforms the other models at a significance level of 0.0124 (NB), 0.0188 (LTM), and 0.0022 (CL). The difference is insignificant compared to ACMNs (0.6384) and SPNs (0.7566).

To compute the probability of one example, MEVMs require as many steps as there are blocks of exchangeable variables. Hence, EM for MEVM is significantly more efficient than EM for NB, both for a single EM iteration and to reach the stopping criterion. While the difference was less significant for problems with fewer than 100 variables, the EM algorithm for MEVM was up to *two orders of magnitude* faster for data sets with 100 or more variables.

6. Discussion

Exchangeable variable models (EVMs) provide a framework for probabilistic models combining the notions of conditional independence and partial exchangeability. As a result, it is possible to efficiently learn the parameters and structure of tractable high tree-width models. EVMs can model complex positive and negative correlations be-

tween large numbers of variables. We presented the theory of EVMs and showed that a particular subfamily is optimal for several important classes of Boolean functions. Experiments with a large number of data sets verified that mixtures of EVMs are powerful and highly efficient models for classification and probability estimation.

EVMs are potential components in deep architectures such as sum-product networks (Gens & Domingos, 2013). In light of Theorem 4.4, exchangeable variable nodes, complementing sum and product nodes, can lead to more compact representations with fewer parameters to learn. EVMs are also related to graphical modeling with perfect graphs (Jebara, 2013). In addition, EVMs provide an insightful connection to lifted probabilistic inference (Kersting, 2012), an active research area concerned with exploiting symmetries for more efficient probabilistic inference. We have developed a principled framework based on partial exchangeability as an important notion of structural symmetry. There are numerous opportunities for cross-fertilization between EVMs, perfect graphical models, collective graphical models, and statistical relational models.

Directions for future work include more sophisticated structure learning, EVMs with continuous variables, EVMs based on instances of partial exchangeability other than finite exchangeability, novel statistical relational formalisms incorporating EVMs, applications of EVMs, and a general theory of graphical models with exchangeable potentials.

Acknowledgments

References

Bui, Hung B., Huynh, Tuyen N., and de Salvo Braz, Rodrigo. Exact lifted inference with distinct soft evidence on every object. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*, 2012.

Canteaut, Anne and Videau, Marion. Symmetric boolean functions. *Information Theory*, 51(8):2791–2811, 2005.

Choi, Myung Jin, Tan, Vincent Y. F., Anandkumar, Animashree, and Willsky, Alan S. Learning latent tree graphical models. *J. Mach. Learn. Res.*, 12:1771–1812, 2011.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theor.*, 14(3): 462–467, 2006.

de Finetti, Bruno. Sur la condition d'équivalence partielle. In *Colloque consacré a la theorie des probabilités*, volume VI, pp. 5–18. Hermann, Paris, 1938. English translation in R. Jeffrey (ed.), pp. 193–205.

Diaconis, Persi and Freedman, David. De Finetti's generalizations of exchangeability. In *Studies in Inductive Logic and Probability*, volume II. 1980a.

Diaconis, Persi and Freedman, David. Finite exchangeable sequences. *The Annals of Probability*, 8(4):745–764, 1980b.

Domingos, Pedro and Pazzani, Michael J. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

Friedman, Nir. The bayesian structural em algorithm. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 129–138, 1998.

Gens, Robert and Domingos, Pedro. Learning the structure of sum-product networks. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 873–880, 2013.

Getoor, Lise and Taskar, Ben. *Introduction to Statistical Relational Learning*. The MIT Press, 2007.

Gupta, Rahul, Diwan, Ajit A., and Sarawagi, Sunita. Efficient inference with cardinality-based clique potentials. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 329–336, 2007.

Jebara, Tony. Perfect graphs and graphical modeling. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2013.

Jennrich, Robert I. and Schluchter, Mark D. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820, 1986.

Kersting, Kristian. Lifted probabilistic inference. In *Proceedings of 20th European Conference on Artificial Intelligence (ECAI)*, pp. 33–38, 2012.

Koller, Daphne and Friedman, Nir. *Probabilistic Graphical Models*. The MIT Press, 2009.

Lauritzen, Steffen L., Barndorff-Nielsen, Ole E., Dawid, A. P., Diaconis, Persi, and Johansen, Søren. Extreme point models in statistics. *Scandinavian Journal of Statistics*, 11(2), 1984.

Lowd, Daniel and Domingos, Pedro. Naive bayes models for probability estimation. In *Proceedings of the 22nd International Conference on Machine learning (ICML)*, pp. 529–536, 2005.

Lowd, Daniel and Rooshenas, Amirmohammad. Learning markov networks with arithmetic circuits. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 406–414, 2013.

Metsis, Vangelis, Androutsopoulos, Ion, and Paliouras, Georgios. Spam filtering with naive bayes - which naive bayes? In *Conference on Email and Anti-Spam (CEAS)*, 2006.

Niepert, Mathias. Markov chains on orbits of permutation groups. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 624–633, 2012.

Niepert, Mathias and Van den Broeck, Guy. Tractability through exchangeability: A new perspective on efficient probabilistic inference. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

Pang, Bo and Lee, Lillian. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, pp. 271–278, 2004.

Rennie, Jason, Shih, Lawrence, Teevan, Jaime, and Karger, David. Tackling the poor assumptions of naive bayes text classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 616–623, 2003.

Rumelhart, David E., Hinton, Geoffrey E., and Williams, Ronald J. Learning internal representations by error propagation. pp. 318–362. MIT Press, 1986.

Sheldon, Daniel and Dietterich, Thomas. Collective graphical models. In *Proceedings of the 25th Conference on Neural Information Processing Systems (NIPS)*, pp. 1161–1169. 2011.

Stefanescu, Catalina and Turnbull, Bruce W. Likelihood inference for exchangeable binary data with varying cluster sizes. *Biometrics*, 59(1):18–24, 2003.

Tarlow, Daniel, Givoni, Inmar E., and Zemel, Richard S. Hopmap: Efficient message passing with high order potentials. In *Proceedings of 13th Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 812–819, 2010.

Tarlow, Daniel, Swersky, Kevin, Zemel, Richard S, Adams, Ryan P, and Frey, Brendan J. Fast exact inference for recursive cardinality models. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 825–834, 2012.

Van Haaren, Jan and Davis, Jesse. Markov network structure learning: A randomized feature generation approach. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 1148–1154, 2012.

A. Proof of Proposition 3.1

Let $\mathbf{X}$ be partially exchangeable with respect to the statistic T with values $\mathcal{T}$, let $|\mathcal{T}| = \mathbf{poly}(|\mathbf{X}|)$, and let, for any partial assignment $\mathbf{e}$, $S_{t,\mathbf{e}} := \{\mathbf{x} \mid T(\mathbf{x}) = t \text{ and } \mathbf{x} \sim \mathbf{e}\}$, where $\mathbf{x} \sim \mathbf{e}$ denotes that $\mathbf{x}$ and $\mathbf{e}$ agree on the variables in their intersection (Koller & Friedman, 2009). If we can in time $\mathbf{poly}(|\mathbf{X}|)$,

(1) for every $\mathbf{e}$ and every $t \in \mathcal{T}$, decide if there exists an $\mathbf{x} \in S_{t,\mathbf{e}}$ and, if so, construct such an $\mathbf{x}$,

then the complexity of MAP inference, that is, computing $\text{argmax}_{\mathbf{y}} P(\mathbf{y}, \mathbf{e})$ for any partial assignment $\mathbf{e}$, is $\mathbf{poly}(|\mathbf{X}|)$. If, in addition, we can in time $\mathbf{poly}(|\mathbf{X}|)$,

(2) for every $\mathbf{e}$ and every $t \in \mathcal{T}$, compute $|S_{t,\mathbf{e}}|$,

then the complexity of marginal inference, that is, computing $P(\mathbf{e})$ for any partial assignment $\mathbf{e}$, is $\mathbf{poly}(|\mathbf{X}|)$.

Proof. We first prove statement (1). Let $\mathbf{e}$ be a given partial assignment and assume we want to compute $\text{argmax}_{\mathbf{y}} P(\mathbf{y}, \mathbf{e})$. We construct an $\mathbf{x}_t \in S_{t,\mathbf{e}}$ for each $t \in \mathcal{T}$ and set $\hat{\mathbf{x}}_t := \text{argmax}_{\mathbf{x}_t} P(\mathbf{x}_t)$. By assumption, this is possible in time $\mathbf{poly}(|\mathbf{X}|)$. Since we have that $\hat{\mathbf{x}}_t = \hat{\mathbf{y}}\mathbf{e}$ with $\hat{\mathbf{y}} := \text{argmax}_{\mathbf{y}} P(\mathbf{y}, \mathbf{e})$ we can extract the solution in linear time.

To prove statement (2), let $\mathbf{e}$ be a partial assignment. We construct a $\mathbf{x}_t \in S_{t,\mathbf{e}}$ for each $t \in \mathcal{T}$ for which such an $\mathbf{x}_t$ exists, compute $|S_{t,\mathbf{e}}|$, and return $\sum_{t \in \mathcal{T}} P(\mathbf{x}_t)|S_{t,\mathbf{e}}|$. By assumption, this is possible in time $\mathbf{poly}(|\mathbf{X}|)$. $\square$

We can utilize Proposition 3.1 to prove that probabilistic inference for a sequence of n exchangeable binary variables is tractable.

Example A.1 (Finite Exchangeability). Let $\mathbf{X}$ be an exchangeable sequence of binary random variables. Let $\mathbf{n}(\mathbf{e})$ be the number of 1s in a partial assignment $\mathbf{e}$ to the variables $\mathbf{X}$. Clearly, we have that $\mathbf{X}$ is exchangeable with respect to the statistic $T(\mathbf{x}) = \mathbf{n}(\mathbf{x})$ with values $\mathcal{T} = \{0, ..., n\}$.

First, we prove that for every partial assignment $\mathbf{e}$ to k of the n variables and every $t \in \mathcal{T}$, we can decide if there exists an $\mathbf{x} \in S_{t,\mathbf{e}}$ and, if so, construct such an $\mathbf{x}$ in time $\mathbf{poly}(|\mathbf{X}|)$. If $\mathbf{n}(\mathbf{e}) > t$ or $n - k + \mathbf{n}(\mathbf{e}) < t$, then there does *not* exist such an $\mathbf{x}$. Otherwise it is possible to generate a $\mathbf{x}$ with $\mathbf{n}(\mathbf{x}) = t$ in linear time by assigning exactly $t - \mathbf{n}(\mathbf{e})$ ones to the unassigned variables and we have that $\mathbf{x} \in S_{t,\mathbf{e}}$. Hence, MAP inference is tractable.

Next, we prove that for every partial assignment $\mathbf{e}$ to k variables and every $t \in \mathcal{T}$, we can compute $|S_{t,\mathbf{e}}|$ in time

$\mathbf{poly}(|\mathbf{X}|)$. But this is possible since $|S_{t,\mathbf{e}}| = \binom{n-k}{t-\mathbf{n}(\mathbf{e})}$. Hence, marginal inference is tractable.

Please note that Example A.1 implies tractability results for numerous important special cases of finite exchangeability such as parity and threshold functions.

There are forms of finite partial exchangeability (Diaconis & Freedman, 1980a) that go beyond the notion of *full* finite exchengeability and, therefore, cardinality-based potentials (Gupta et al., 2007; Tarlow et al., 2010) of Example A.1. We provide three examples.

Example A.2 (Block Exchangeability). Let w be a fixed constant. For a sequence of binary random variables $\mathbf{X}$ let $\mathcal{X} = \{\mathbf{X}_1, ..., \mathbf{X}_w\}$ be a partition of the variables $\mathbf{X}$ into w subsequences, and let $\mathbf{n}_{\mathbf{Y}}(\mathbf{x})$ be the number of 1s in an assignment $\mathbf{x}$ projected onto the variables $\mathbf{Y} \subseteq \mathbf{X}$. Now, let $T(\mathbf{x}) = (\mathbf{n}_{\mathbf{X}_1}(\mathbf{x}), ..., \mathbf{n}_{\mathbf{X}_w}(\mathbf{x}))$.

It is straight-forward to verify that $|\mathcal{T}| = \mathbf{poly}(|\mathbf{X}|)$. Moreover, with arguments similar to those made in Example A.1 one can show that conditions (1) and (2) of Proposition 3.1 are met. Hence, MAP and marginal inference are tractable for the statistic T.

Example A.3. Let $\mathbf{X}$ be a sequence of n binary random variables and let $\tau_{0 \to 1}(\mathbf{x})$ be the number of times 01 occurs as a substring[2] in $\mathbf{x}$. Now, consider the statistic

$$T(\mathbf{x}) = \tau_{0 \to 1}(\mathbf{x}).$$

For example, for $\mathbf{x} = 11011111$ we have $T(\mathbf{x}) = 1$ and for $\mathbf{x} = 01010101$ we have $T(\mathbf{x}) = 4$. We also have that $|\mathcal{T}| = \lfloor n/2 \rfloor + 1 = \mathbf{poly}(|\mathbf{X}|)$.

Now, let $\mathbf{e}$ be a partial assignment to k of the n variables and let $0 \le t \le \lfloor n/2 \rfloor$ be a value of the statistic. Let $\mathbf{b} = \{0, 1, *\}^n$ be a string where the characters 0 and 1 encode the assignments to variables according to $\mathbf{e}$ and the character * encodes unassigned variables. We now partition $\mathbf{b}$ into four sets G_{ij}, $i, j \in \{0, 1\}$, of substrings defined as $G_{ij} := \{\mathbf{s} \sqsubseteq \mathbf{b} \mid s_1 = i, s_{|\mathbf{s}|} = j, s_\ell = * \text{ for } 1 \le i < \ell < j \le |\mathbf{s}|\}$, where $\sqsubseteq$ denotes the substring relation. We can now complete the partial assignment $\mathbf{e}$ to a joint assignment $\mathbf{x}$ with $T(\mathbf{x}) = t$ if and only if (1) $\tau_{0 \to 1}(\mathbf{b}) + |G_{01}| \le t$ and (2) $\tau_{0 \to 1}(\mathbf{b}) + \sum_{\mathbf{s} \in G_{00}} \left\lceil \frac{|\mathbf{s}| - 2}{2} \right\rceil + \sum_{\mathbf{s} \in G_{01}} \left\lfloor \frac{|\mathbf{s}|}{2} \right\rfloor + \sum_{\mathbf{s} \in G_{10}} \left\lfloor \frac{|\mathbf{s}| - 2}{2} \right\rfloor + \sum_{\mathbf{s} \in G_{11}} \left\lceil \frac{|\mathbf{s}| - 2}{2} \right\rceil \ge t$. When these two conditions are met, the full assignment $\mathbf{x}$ can be constructed by completing the substring in the groups G_{ij} so as to make $T(\mathbf{x}) = t$ and this is possible in linear time. Hence, MAP inference is tractable.

It is possible to construct novel tractable statistics by nesting statistics that are known to be tractable.

---

[2]As opposed to subsequences, substrings are consecutive parts of a string.

Example A.4 (Nested Tractable Statistics). Let $\mathbf{X}$ be an $n \times n$ array of binary random variables. For instance, $\mathbf{X}$ could represent a binarized image with n rows and n columns. Let k be a fixed integer constant and let ℓ be the integer such that $n = k\ell$. We assume without loss of generality that such an integer exists. We partition the original array into ℓ^2 squares of dimension $k \times k$. For $1 \leq i \leq \ell^2$, let $\mathbf{S}_i$ be the variables of square i. Now, let $T_1 : \{0,1\}^{k^2} \to \{0,1\}$ be the statistic defined as

$$T_1(\mathbf{s} = (s_1, ..., s_{k^2})) = [[\sum_{i=1}^{k^2} s_i > \tau]],$$

for some τ with $0 \leq \tau < k^2$. That is, $T_1(\mathbf{s}) = 1$, if the number of 1s in a given square exceeds a threshold of τ and $T_1(\mathbf{s}) = 0$ otherwise. Please note that for $\tau = 0$ this corresponds to max-pooling. Now, let $T : \{0,1\}^{n^2} \to \{0, ..., \ell^2\}$ be the statistic defined as follows:

$$T(\mathbf{x}) = \sum_{i=1}^{\ell^2} T_1(\mathbf{s}_i).$$

Based on the tractability of the two statistics, it is straight-forward to verify that both MAP and marginal inference is tractable for the statistic T.

Please note that the presented theoretical framework facilitates the discovery and development of novel tractable non-local potentials.

B. Proof of Theorem 3.2

Let $X_1, ..., X_n$ be a sequence of random variables with joint distribution P, let T be a statistic with distinct values $t_0, ..., t_k$, and let $X_1, ..., X_n$ be partially exchangeable with respect to T. The ML estimates for N examples, $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$, are $\mathtt{MLE}[(w_0, ..., w_k)] = \left(\frac{c_0}{N}, ..., \frac{c_k}{N}\right)$, where $c_i = \sum_{j=1}^{N}[[T\left(\mathbf{x}^{(j)}\right) = t_i]]$.

Proof. Let $\theta = (w_0, ..., w_k)$. By Theorem 2.3, the log-likelihood for N examples $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$ is

$$\mathtt{L}(\theta) = \sum_{j=1}^{N} \log \left(\sum_{i=0}^{k} w_i U_i \left(\mathbf{x}^{(j)}\right) \right).$$

Let $c_i = \sum_{j=1}^{N}[[T\left(\mathbf{x}^{(j)}\right) = t_i]]$ and let $\hat{\mathbf{x}}_i$ be a joint assignment with $T(\hat{\mathbf{x}}_i) = t_i$. Then, $\mathtt{L}(\theta) = \sum_{i=0}^{k} c_i \log(w_i U_i(\hat{\mathbf{x}}_i)) = \sum_{i=0}^{k} c_i[\log(w_i) + \log(U_i(\hat{\mathbf{x}}_i))] = \sum_{i=0}^{k} c_i \log(w_i) + \sum_{i=0}^{k} c_i \log(U_i(\hat{\mathbf{x}}_i))$. The second term is free of parameters and, hence, finding the ML estimates amounts to maximizing the first sum. This is equivalent to finding the maximum likelihood estimate of a multinomial which can be solved with Lagrange multipliers. Hence, $\mathtt{MLE}(w_i) = \frac{c_i}{N}$, for $0 \leq i \leq k$. $\square$

C. Proof of Proposition 3.3

The following statements are necessary conditions for exchangeability of a finite sequence of random variables $X_1, ..., X_n$. For all $i, j, i', j' \in \{1, ..., n\}$ with $i \neq j$ and $i' \neq j'$

(1) $\mathbf{E}(X_i) = \mathbf{E}(X_j)$;

(2) $\mathbf{Var}(X_i) = \mathbf{Var}(X_j)$; and

(3) $\mathbf{Cov}(X_i, X_j) = \mathbf{Cov}(X_{i'}, X_{j'}) \geq -\frac{\mathbf{Var}(X_i)}{(n-1)}$.

These conditions are well-known and are straight-forward to prove. Nevertheless, for the sake of completeness, we prove statement (3).

Proof. It is straight-forward to prove statements (1) and (2). In order to prove statement (3) we use statements (2) to write

$$0 \leq \mathbf{Var}(X_1 + \cdots + X_n)$$
$$= \mathbf{Var}(X_1) + \cdots + \mathbf{Var}(X_n) + 2\sum_{i<j} \mathbf{Cov}(X_i, X_j)$$
$$= n\mathbf{Var}(X_i) + n(n-1)\mathbf{Cov}(X_i, X_j).$$

Hence, $\mathbf{Cov}(X_i, X_j) \geq -\frac{\mathbf{Var}(X_i)}{(n-1)}$. $\square$

D. Proof of Theorem 4.4

The mixtures of EVMs family is globally optimal under zero-one loss for

1. Conjunctions and disjunctions of attributes;

2. Symmetric Boolean functions such as

   - Threshold (m-of-n) functions
   - Parity functions
   - Counting functions
   - Exact value functions

Proof. Let $\mathbf{X}$ be the sequence of variables under consideration. We write $y(\mathbf{x})$ for the (hidden) class value of example $\mathbf{x}$. For conjunctions of attributes, let $\hat{\mathbf{X}} \subseteq \mathbf{X}$ be the sequence of variables that are part of the conjunction. Conditioned on the binary class variable being either 0 or 1, we partition the variables into the two blocks $\hat{\mathbf{X}}$ and $\mathbf{X} - \hat{\mathbf{X}}$. We set the parameters of the MEVM as follows.

$q_{(\hat{\mathbf{X}})}(\ell \mid 1) = 1.0$ if $\ell = |\hat{\mathbf{X}}|$ and $q_{(\hat{\mathbf{X}})}(\ell \mid 1) = 0.0$ otherwise;

$q_{(\hat{\mathbf{X}})}(\ell \mid 0) = 0.0$ if $\ell = |\hat{\mathbf{X}}|$ and $q_{(\hat{\mathbf{X}})}(\ell \mid 0) = \frac{\binom{|\hat{\mathbf{X}}|}{\ell}}{2^{|\hat{\mathbf{X}}|}}$ otherwise;

$$q_{(\mathbf{X}-\hat{\mathbf{X}})}(\ell \mid 1) = \frac{\binom{|\mathbf{X}|-|\hat{\mathbf{X}}|}{\ell}}{2^{|\mathbf{X}|-|\hat{\mathbf{X}}|}}; \quad q_{(\mathbf{X}-\hat{\mathbf{X}})}(\ell \mid 0) = \frac{\binom{|\mathbf{X}|-|\hat{\mathbf{X}}|}{\ell}}{2^{|\mathbf{X}|-|\hat{\mathbf{X}}|}};$$

$$p(1) = \frac{2^{|\mathbf{X}|-|\hat{\mathbf{X}}|}}{2^{|\mathbf{X}|}}; \text{ and } p(0) = \frac{(2^{|\hat{\mathbf{X}}|}-1)(2^{|\mathbf{X}|-|\hat{\mathbf{X}}|})}{2^{|\mathbf{X}|}}.$$

Then, we have that $\mathbf{P}(1 \mid \mathbf{x}) > 0$ if $y(\mathbf{x}) = 1$ and $\mathbf{P}(1 \mid \mathbf{x}) = 0$ otherwise. Moreover, $\mathbf{P}(0 \mid \mathbf{x}) = 0$ if $y(\mathbf{x}) = 1$ and $\mathbf{P}(0 \mid \mathbf{x}) > 0$ otherwise. Hence, the MEVM classifier always returns the correct class value. A similar argument can be made to prove the optimality for disjunctions of attributes.

To prove the second statement, we consider an MEVM model with a binary class variable and the following block structure. For each of the class variable's values y, $y \in \{0,1\}$, we have that $\mathcal{X}_y = \{X_1, ..., X_n\}$. That is, conditioned on each class value, the attributes are assumed to be exchangeable (see Figure 3; right). It is straightforward to verify that this particular MEVM can learn *arbitrary* discrete distributions over any symmetric Boolean function. □