International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Hierarchical models as marginals of hierarchical models

Guido Montúfar [a,*], Johannes Rauh [b]

[a] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany
[b] Department of Mathematics and Statistics, York University, 4700 Keele St., Toronto, ON, M3J 1P3, Canada

ARTICLE INFO

Article history:
Received 22 March 2016
Received in revised form 20 August 2016
Accepted 6 September 2016
Available online xxxx

Keywords:
Hierarchical model
Restricted Boltzmann machine
Soft-plus unit
Rectified linear unit
Interaction model
Graphical model

ABSTRACT

We investigate the representation of hierarchical models in terms of marginals of other hierarchical models with smaller interactions. We focus on binary variables and marginals of pairwise interaction models whose hidden variables are conditionally independent given the visible variables. In this case the problem is equivalent to the representation of linear subspaces of polynomials by feedforward neural networks with soft-plus computational units. We show that every hidden variable can freely model multiple interactions among the visible variables, which allows us to generalize and improve previous results. In particular, we show that a restricted Boltzmann machine with $[2(\log(v)+1)/(v+1)]2^v-1$ hidden binary variables can approximate every distribution of v visible binary variables arbitrarily well, which improves the previous bound $2^{v-1}-1$.

1. Introduction

Consider a finite set V of random variables. A hierarchical log-linear model is a set of joint probability distributions that can be written as products of interaction potentials, as $p(x)=\prod_\Lambda \psi_\Lambda(x)$, where $\psi_\Lambda(x)=\psi_\Lambda(x_\Lambda)$ only depends on the subset Λ of variables and where the product runs over a fixed family of such subsets. By introducing hidden variables, it is possible to express the same probability distributions in terms of potentials which involve only small sets of variables, as $p(x)=\sum_y \prod_\lambda \psi_\lambda(x,y)$, with small sets λ. Using small interactions is a central idea in the context of connectionist models [18], where the sets λ are restricted to have cardinality two. Models with small interaction sets are particularly well suited for Gibbs sampling [6]. The representation, or explanation, of complex interactions among observed variables in terms of hidden variables relates to a variety of problems in deep learning [7] and the study of common ancestors [19].

We are interested in sufficient and necessary conditions on the number of hidden variables, their values, and the interaction structures, under which the visible marginals are flexible enough to represent any distribution from a given hierarchical model. Many problems can be formulated as special cases of this general problem.

In this article, we focus on the case that all variables are binary. For the hierarchical models with hidden variables, we restrict our attention to models involving only pairwise interactions and whose hidden variables are conditionally independent given the visible variables (no direct interactions between the hidden variables). An important example of this type of models is the restricted Boltzmann machine (RBM), which has full bipartite interactions between the visible and hidden variables. The representational power of RBMs has been studied in many papers; see, e.g., [5,20,9,11]. The free energy function of such a model is a sum of soft-plus computational units $x\mapsto \log(1+\exp(\sum_{i\in V} w_i x_i + c))$. On the other hand,

* Corresponding author.
E-mail addresses: montufar@mis.mpg.de (G. Montúfar), jarauh@yorku.ca (J. Rauh).
URL: http://personal-homepages.mis.mpg.de/montufar/ (G. Montúfar).

http://dx.doi.org/10.1016/j.ijar.2016.09.003
0888-613X/

the energy function of a fully observable hierarchical model with binary variables is a polynomial, with monomials corresponding to pure interactions. Since any function of binary variables can be expressed as a polynomial, the task is then to characterize the polynomials computable by soft-plus units with binary inputs. Our analysis of soft-plus units also allows us to describe the polynomials computable by rectified linear units $x \mapsto \max\{0, \sum_{i\in V} w_i x_i + c\}$, which can be regarded as limits of soft-plus units with large magnitude parameters.

Younes [20] showed that a hierarchical model with v binary variables and a total of M pure higher order interactions (among three or more variables) can be represented as the visible marginal of a pairwise interaction model with M hidden binary variables. In Younes' construction, each pure interaction is modeled by one hidden binary variable that interacts pairwise with each of the involved visible variables. In fact, he shows that this replacement can be accomplished without increasing the number of model parameters, by imposing linear constraints on the coupling strengths of the hidden variable. In this work we investigate how to squeeze more degrees of freedom out of each hidden variable. An indication that this should be possible is the fact that the full interaction model (which contains all strictly positive joint distributions of the visible variables), for which $M = 2^v - \binom{v}{2} - v - 1$, can be modeled by a pairwise interaction model with $2^{v-1} - 1$ hidden binary variables [11]. Indeed, by controlling groups of polynomial coefficients at the time, we show that, in general, less than M hidden variables are sufficient.

Our results lead to new universal approximation results for RBMs. Namely, we show that $[2(\log(v) + 1)/(v + 1)]2^v - 1$ hidden binary variables are sufficient to approximate every distribution of v visible binary variables arbitrarily well (see Theorem 11 and Corollary 12). This upper bound differs only by a logarithmic factor from the hard lower bound $\lceil 2^v/(v + 1)\rceil - 1$ that results from demanding that the RBM has $2^v - 1$ parameters, which is the dimension of the set of distributions of v binary variables. This represents a significant improvement of the previous upper bound $2^{v-1} - 1$ from [11].

A special case of hierarchical models with hidden variables are mixtures of hierarchical models. The smallest mixtures of hierarchical models that contain other hierarchical models have been studied in [10]. Our analysis of soft-plus polynomials is different and complementary to the approach followed there. For the necessary conditions, the idea there is to compare the possible support sets of the limit distributions of both models. For the sufficient conditions, the idea is to find a small S-set covering of the set of elementary events. An S-set of a probability model is a set of elementary events such that every distribution supported in that set is a limit distribution from the model.

This article is organized as follows. Section 2 introduces hierarchical models and formalizes our problem in the light of previous results. Section 3 pursues a characterization of the polynomials that can be represented by soft-plus units. Section 4 applies this characterization to study the representation of hierarchical models in terms of pairwise interaction models with hidden variables. This section addresses mainly restricted Boltzmann machines. Section 5 offers our conclusions and outlook. This article is an extended version of the workshop article [14].

2. Preliminaries

This section introduces hierarchical models, with and without hidden variables, formalizes the problem that we address in this paper, and presents motivating prior results.

2.1. Hierarchical models

Consider a finite set V of variables with finitely many joint states $x = (x_i)_{i\in V} \in \mathbb{X} = \bigtimes_{i\in V} \mathbb{X}_i$. Denote by $v = |V|$ the cardinality of V. For a given set $S \subseteq 2^V$ of subsets of V let

$$\mathcal{V}_{\mathbb{X},S} := \left\{ g(x) = \sum_{\Lambda\in S} g_\Lambda(x) \colon g_\Lambda(x) = g_\Lambda(x_\Lambda) \right\}.$$

This is the linear subspace of $\mathbb{R}^{\mathbb{X}}$ spanned by functions g_Λ that only depend on sets of variables $\Lambda \in S$. For convenience, in all what follows we assume that S is a simplicial complex, meaning that $A \in S$ implies $B \in S$ for all $B \subseteq A$. Furthermore, we assume that the union of elements of S equals V. The function space $\mathcal{V}_{\mathbb{X},S}$ is finite dimensional, and it is not difficult to write down an explicit basis. When all variables are binary, $\mathbb{X}_i = \{0, 1\}$ for all $i \in V$, then a basis of $\mathcal{V}_{\mathbb{X},S}$ is given by the squarefree monomial functions

$$x^\Lambda := \prod_{i\in\Lambda} x_i, \quad \Lambda \in S.$$

The attribute *squarefree* refers to the fact that no variable appears to a power larger than one. The set $\mathcal{V}_{\mathbb{X},S}$ then consists of all polynomials in x_i, $i \in V$, that only involve squarefree monomials x^Λ with $\Lambda \in S$.

The hierarchical model of probability distributions on $\mathbb{X}$ with interactions S is the set

$$\mathcal{E}_{\mathbb{X},S} := \left\{ p(x) = \frac{1}{Z(g)} \exp(g(x)) \colon g \in \mathcal{V}_{\mathbb{X},S} \right\}, \tag{1}$$

where $Z(g) = \sum_{x'\in\mathbb{X}} \exp(g(x'))$ is a normalizing factor. We call

Please cite this article in press as: G. Montúfar, J. Rauh, Hierarchical models as marginals of hierarchical models, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.09.003