Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

Maximal Information Divergence from Statistical Models defined by Neural Networks

by

Guido Montúfar, Johannes Rauh, and Nihat Ay

Preprint no.: 31 2013

Maximal Information Divergence from Statistical Models defined by Neural Networks

Guido Montúfar[1], Johannes Rauh[2], and Nihat Ay[2,3]

[1] Department of Mathematics, Pennsylvania State University,
University Park, PA 16802, USA,
gfm10@psu.edu,
[2] Max Planck Institute for Mathematics in the Sciences,
Inselstraße 22, 04103 Leipzig, Germany,
{jrauh,nay}@mis.mpg.de,
[3] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.

Abstract. We review recent results about the maximal values of the Kullback-Leibler information divergence from statistical models defined by neural networks, including naïve Bayes models, restricted Boltzmann machines, deep belief networks, and various classes of exponential families. We illustrate approaches to compute the maximal divergence from a given model starting from simple sub- or super-models. We give a new result for deep and narrow belief networks with finite-valued units.

Keywords: neural network, exponential family, Kullback-Leibler divergence, multi-information

1 Introduction

In statistical learning theory, probability models are used to infer representations of data. In model selection it is often assumed that the model approximation errors are negligible compared with the statistical approximation errors. This assumption may not always be justified in practice; in some cases even full dimensional models only fill a small portion of the space of probability distributions, and telling the general structure of the data generating distributions, in order to constrain the possible model classes, is difficult.

Here we take a complementary perspective, disregarding the statistical approximation errors and focussing on the model approximation errors. We quantify the model approximation error of a model $\mathcal{M}$ by the divergence function $p \mapsto D(p\|\mathcal{M}) = \inf_{q\in\mathcal{M}} D(p\|q)$, where $D(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)}$ is the Kullback-Leibler divergence from p to q.[4] We study the maximum value of $D(\cdot\|\mathcal{M})$, which corresponds to a worst-case analysis. The ideas from this paper can also be used to study the expectation value given a prior on the set of target distributions, see [16]. The model approximation error can be used as a

[4] We formulate our results in such a way that they are independent from the logarithm's base used in the definition of the divergence.

criterion for model selection. Related ideas are discussed in [2] in the context of model design and reinforcement learning.

Most probability models with hidden variables are singular and not identifiable. Moreover, data distributions that are not contained in these models can have several maximum likelihood estimates. Although controlling parameter-identifiability is crucial when estimating learning coefficients in Bayesian model selection, we will instead focus on the value of the data likelihood and the sets of maximizing distributions, irrespective of their parameters.

In general, the function $D(\cdot\|\mathcal{M})$ has no explicit formula, making the estimation of the maximizers and the maximum value difficult. For exponential families the situation is slightly better, as for each distribution p the divergence $D(p\|\cdot)$ has a unique minimizer over $\overline{\mathcal{M}}$. For certain families, such as independence models and convex exponential families, there even is a closed formula for this function. The approximation properties of various classes of exponential families have been studied in [9,10,1,18,19,16,6]. The divergence from complicated models can be estimated by finding tractable exponential subfamilies. This idea was used in [17] to study approximation errors of restricted Boltzmann machines.

The representational power of neural networks has been studied for many years and by too many authors to refer to appropriately at this place, see for instance [3,5,4]. The representational power of the networks discussed in this paper has been studied, in particular, in [7,20,8,13,17,11].

Section 2 reviews bounds on $D_{\mathcal{M}}$ for statistical models defined by neural networks and for exponential families. Section 3 discusses strategies to bound $D_{\mathcal{M}}$ via sub-models and super-models, and discusses a class of exponential families contained in restricted Boltzmann machines and deep belief networks. Section 4 puts our results in perspective.

2 Maximal information divergence

We consider neural networks with a set of visible units $X_1, \dots, X_n$, where each X_i takes values in a finite set $\mathcal{X}_i$ of cardinality $|\mathcal{X}_i| = N_i$. See Fig. 1. The visible state space of such a system is $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$. For any subset $A \subseteq [n]$ let $N_A = \prod_{i\in A} N_i$ be the number of joint states of the units indexed by A, and let $N = N_{[n]} = |\mathcal{X}|$. We denote the set of all probability distributions $(p_x)_{x\in\mathcal{X}}$ on $\mathcal{X}$ by $\Delta(\mathcal{X})$, or Δ if $\mathcal{X}$ is understood. The ***maximal information divergence*** from a model $\mathcal{M} \subseteq \Delta(\mathcal{X})$ is $D_{\mathcal{M}} := \max_{p\in\Delta} D(p\|\mathcal{M})$. An *$rI$-projection* of $p \in \Delta$ onto $\mathcal{M}$ is a point $p_{\mathcal{M}}$ in the closure $\overline{\mathcal{M}}$ of $\mathcal{M}$ with $D(p\|\mathcal{M}) = D(p\|p_{\mathcal{M}})$.

2.1 Probability models defined by neural networks

The ***independence model*** $\mathcal{E}_n^1$ of n variables $X_1, \dots, X_n$ is the set of probability distributions of the form $p(x) = \prod_{i\in[n]} p_i(x_i)$ for all $x = (x_1, \dots, x_n) \in \mathcal{X}$. This model describes non-interacting stochastic variables. The following result is due to Ay and Knauf [1, Corollary 4.10].

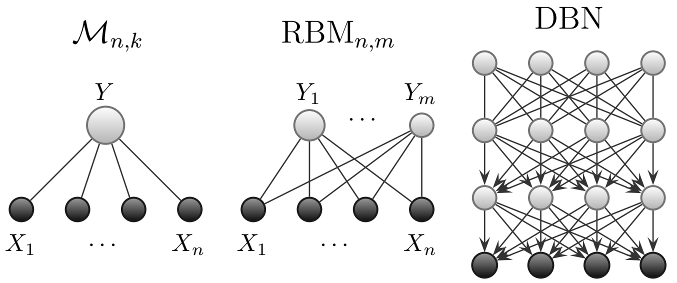

Fig. 1. The naïve Bayes model $\mathcal{M}_{n,k}$, the restricted Boltzmann machine RBM$_{n,m}$, and a deep belief network. Light (dark) nodes represent hidden (visible) variables.

Lemma 1. *The maximal divergence to* $\mathcal{E}_n^1$ *is bounded by*

$$D_{\mathcal{E}_n^1} \leq \log(N/\max_{i\in[n]} N_i) \ .$$

If all variables are q-ary, then $D_{\mathcal{E}_n^1} = (n-1)\log(q)$*, and the maximizers are the uniform distributions on q-ary codes of cardinality* q *and minimum distance* n*.*

The ***mixture of product distributions*** $\mathcal{M}_{n,k}$, or ***naïve Bayes model***, is the graphical model on a star graph, where the leaves are visible variables, and the internal node is a hidden variable with k states.

Theorem 1. *Let* $A \subseteq [n]$*. If* $k \geq N_{[n]\setminus A}$*, then* $D_{\mathcal{M}_{n,k}}$ *is bounded by*

$$D_{\mathcal{M}_{n,k}} \leq \log(N_A/\max_{j\in A} N_j) \ .$$

When all visible variables are binary, we have the tighter bound

$$D_{\mathcal{M}_{n,k}} \leq \big(n - \lfloor\log_2(k)\rfloor - \frac{k}{2^{\lfloor\log_2(k)\rfloor}}\big)\log(2) \ .$$

Note the similarity of the bounds given in Lemma 1 and Theorem 1, In fact Theorem 1 can be derived from Lemma 1, together with Lemma 6 given below.

The ***restricted Boltzmann machine*** RBM$_{n,m}$ is the undirected stochastic network with full bipartite interaction graph $K_{n,m}$, where an independent set of m units is hidden, and an independent set of n units is visible.

Theorem 2. *Let* $A \subseteq [n]$*, and let* $M_1, \dots, M_m$ *be the sizes of the state spaces of the hidden variables. If* $1 + \sum_{j\in[m]}(M_j - 1) \geq N_{[n]\setminus A}$*, then*

$$D_{\mathrm{RBM}_{n,m}} \leq \log(N_A/\max_{j\in A} N_j) \ .$$

When all units are binary, and $m \leq 2^{n-1} - 1$*, we have the tighter bound*

$$D_{\mathrm{RBM}_{n,m}} \leq \big(n - \lfloor\log_2(m+1)\rfloor - \frac{m+1}{2^{\lfloor\log_2(m+1)\rfloor}}\big)\log(2) \ .$$

Theorem 2 subsumes divergence bounds for naïve Bayes models (when $m = 1$) and independence models (when $m = 0$). This result was shown in the binary case in [17, Theorem 2] and in the non-binary case in [15, Theorem 29].

A ***deep belief network*** (DBN) is a layered stochastic network with undirected bipartite interactions between the units in the deepest two layers, which form an RBM, and directed bipartite interactions between all other pairs of subsequent layers, directed towards the first layer, which is the only visible layer.

Theorem 3. *Consider a DBN with L layers, each layer containing n units with state spaces of cardinalities $q_1, \dots, q_n$. Let m be any integer with $\prod_{j=m+2}^{n} q_j \leq m \leq n$, and let $q_1 \geq \cdots \geq q_m$. If $L \geq 2 + \frac{q_1^S - 1}{q_1 - 1}$ for some $S \in \{0, 1, \dots, m\}$, then*

$$D_{\mathrm{DBN}} \leq \log(N_{[m-S]}) \ .$$

In particular, when all units are binary and the network has $L \geq 1 + 2^S$ layers of size $n = 2^{k-1} + k$, for some $S \in \{0, 1, \dots, 2^{k-1}\}$, then

$$D_{\mathrm{DBN}} \leq \left(2^{k-1} - S\right) \log(2) \ .$$

The binary case is [13, Theorem 2], together with [14, Theorem 18]. The non-binary case is new (details in [12]).

The bounds in Theorems 1, 2 and 3 vanish when the number of hidden units is large enough (depending on their state spaces). In this case, the models can approximate all probability distributions on the states of their visible units arbitrarily well, i.e., they are *universal approximators.*

All these theorems can be proved using the same strategy: First, a family of exponential sub-models is identified, and then, the divergence from the union of these sub-models is bounded from above, as in Theorem 8 below.

2.2 Exponential families

Exponential families are widely-used statistical models. Examples include log-linear models, hierarchical models, and independence models. The information divergence maximization problem is by far better understood for exponential families than for other probability models. We use exponential families to approximate probability models with hidden variables.

Let $\varrho = \{A_1, \dots, A_m\}$ be a partition of $\mathcal{X}$. The ***partition model*** $\mathcal{P}_\varrho$ consists of all $p \in \Delta$ with $p(x) = p(y)$ whenever x, y belong to the same block of ϱ; that is, the conditional distribution of p, conditioned on any block $A_i \in \varrho$, equals the uniform distribution. Partition models are the convex exponential families that contain the uniform distribution. The following is a special case of [9, Corollary 1]:

Lemma 2. *Let $\varrho = \{A_1, \dots, A_m\}$ be a partition of $\mathcal{X}$, and denote by $c(\varrho) = \max_{i \in [m]} |A_i|$ the coarseness of ϱ. Then $D_{\mathcal{P}_\varrho} = \log(c(\varrho))$, and the global maximizers are the distributions p with $\mathrm{supp}(p) \cap A_i \leq 1$ for all $i \in [m]$, where $\mathrm{supp}(p) \cap A_i = 1$ holds only if $|A_i| = c(\varrho)$.*

The maximal divergence from any exponential family of dimension k can be bounded from below as follows, see [19, Theorem 28]:

Theorem 4. *Let $\mathcal{E}$ be an exponential family of dimension k. Then*

$$D_{\mathcal{E}} \geq \log(N) - \log(k+1) \ .$$

If equality holds, then $\mathcal{E}$ is a partition model with homogeneous partition.

Probability models defined as marginals of exponential families can behave very different from proper exponential families. Any finite subset of Δ can be embedded in a projection of a two-dimensional exponential family, see [2]:

Lemma 3. *Given any finite set of probability distributions $\{p^{(i)}\}_{i=1}^{K} \subset \Delta_{N-1}$, there is a two-dimensional exponential family $\mathcal{E} \subseteq \Delta_{K-1}$, and a linear map $\psi\colon \Delta_{K-1} \to \Delta_{N-1}$, such that $\psi(\overline{\mathcal{E}}) \supseteq \{p^{(i)}\}_{i=1}^{K}$.*

3 Estimating the information divergence

3.1 Subfamilies and superfamilies

If $\mathcal{M}' \subseteq \mathcal{M}$ then $D_{\mathcal{M}} \leq D_{\mathcal{M}'}$. In special cases it is possible to have equality.

Lemma 4. *If $\mathcal{M}' \subseteq \mathcal{M}$ and if p is a maximizer of the divergence from $\mathcal{M}$ such that $\mathcal{M}'$ contains an rI-projection $p_{\mathcal{M}}$ of p to $\mathcal{M}$, then p maximizes the divergence from $\mathcal{M}'$ among the set $\{q \in \Delta : q_{\mathcal{M}} \in \mathcal{M}' \text{ for some } rI\text{-projection } q_{\mathcal{M}}\}$.*

The lemma ist useful for exponential families; where the set of distributions whose rI-project to $\mathcal{M}$ lies in $\mathcal{M}'$, can be parametrized via $\mathcal{M}' + \mathcal{N}$, where $\mathcal{N}$ is the normal space of $\mathcal{M}$. The following argument due to Juríček [6] is an example:

Let $\mathcal{M} = \mathcal{E}_n^1$ be the independence model of n q-ary variables and let $\mathcal{M}'$ be the set of i.i.d. distributions. By Lemma 1, the uniform distribution p on the states $(1,\dots,1),(2,\dots,2),\dots,(q,\dots,q)$ maximizes the divergence from $\mathcal{M}$, and it is exchangeable. Since the rI-projections of the set of exchangeable distributions to $\mathcal{M}$ belong to $\mathcal{M}'$, Lemma 4 implies that p maximizes the divergence from $\mathcal{M}'$ among the exchangeable distributions, with divergence $D(p\|\mathcal{M}') = (n-1)\log(q)$. Now, $\mathcal{M}'$ as a subset of the exchangeable simplex can be identified with the multinomial model. This proves the following result [6, Theorem 1.1].

Theorem 5. *The maximal divergence from the multinomial model of n q-ary variables is equal to $(n-1)\log(q)$.*

Conversely, simple subfamilies can be used to study larger models:

Lemma 5. *Let $\mathcal{E}$ be an exponential family. Let $\mathcal{M}_i$ be a sub-model of $\mathcal{E}$ with $D_{\mathcal{M}_i} = K$ and divergence maximizers $\mathcal{G}_i$, for all $i \in [k]$. If there is a point $p \in \mathcal{G} = \cap_i \mathcal{G}_i$ with $p_{\mathcal{E}} \in \cup_i \mathcal{M}_i$, then $D_{\mathcal{E}} = K$ and the divergence maximizers are exactly the points in $\mathcal{G}$ whose rI-projections onto $\mathcal{E}$ lie in $\cap_i \mathcal{M}_i$.*

Lemma 5 can be used to prove the homogeneous case of Lemma 1 as follows: The independence model of n q-ary variables contains the partition model $\mathcal{P}_i$ with partition blocks $\{x\colon x_i = y_i\}$ for all $y_i \in \mathcal{X}_i$, for any $i \in [n]$. By Lemma 2, the maximal divergence from the partition model $\mathcal{P}_i$ is $D_{\mathcal{P}_i} = (n-1)\log(q)$, and the set of maximizers is the set $\mathcal{G}_i$ of distributions p whose support $\operatorname{supp}(p) = \{x^{(j)}\}_j$ satisfies $x_i^{(j)} \neq x_i^{(j')}$ for all $j \neq j'$. The intersection $\mathcal{G} = \cap_i \mathcal{G}_i$ is the set of probability distributions with support on a code of minimum distance n. The rI-projection of an arbitrary element $p \in \mathcal{G}$ lies in $\cap_i \mathcal{P}_i = \{u\}$ if and only if p is a uniform distribution on a code of minimum distance n and cardinality q. By Lemma 5 these are the global divergence maximizers from $\mathcal{E}_n^1$.

3.2 Mixtures of exponential families with disjoint supports

The *mixture* $\operatorname{Mixt}(\mathcal{M}_1, \dots, \mathcal{M}_k)$ of k models $\mathcal{M}_1, \dots, \mathcal{M}_k \subseteq \Delta$ is the set of probability distributions of the form $p = \sum_{i=1}^k \lambda_i p^{(i)}$, where $\lambda \in \Delta_{k-1}$ and $p^{(i)} \in \mathcal{M}_i$ for all $i \in [k]$. In general, mixtures are difficult to describe, even for simple models $\mathcal{M}_1, \dots, \mathcal{M}_k$. The situation is much simpler when mixing models supported on disjoint subsets of $\mathcal{X}$:

Lemma 6. *Let $\{A_1, \dots, A_k\}$ be a partition of $\mathcal{X}$ and let $\mathcal{M}_1, \dots, \mathcal{M}_k$ be statistical models with $\mathcal{M}_i \subseteq \Delta(A_i)$. For any $p \in \Delta(\mathcal{X})$, the rI-projections of p to* $\operatorname{Mixt}(\mathcal{M}_1, \dots, \mathcal{M}_k)$ *are the distributions of the form*

$$p_{\mathcal{M}}(x) = p(A_i) p_{\mathcal{M}_i}(x), \qquad \textit{for all } x \in A_i \quad \textit{for all } i \in [k],$$

where $p_{\mathcal{M}_i}$ denotes an rI-projection of $p(x|A_i)$ to $\mathcal{M}_i$ for all $i \in [k]$.

We call a set $\mathcal{Y} \subseteq \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ *cubical* if it can be written as a product $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$ with $\mathcal{Y}_i \subseteq \mathcal{X}_i$ for all $i \in [n]$. A set $\mathcal{Y}$ is cubical iff there exists a product distribution p with $\operatorname{supp}(p) := \{x \in \mathcal{X} : p(x) > 0\} = \mathcal{Y}$ (in this case $\mathcal{Y}_i = \operatorname{supp}(p_i)$). We call a partition *cubical* if it consists of cubical blocks. For any cubical set $\mathcal{Y}$ let $\overline{\mathcal{E}_{\mathcal{Y}}^1}$ denote the set of product distributions with support $\mathcal{Y}$.

Let $\varrho = \{A_1, \dots, A_k\}$ be a cubical partition of $\mathcal{X}$. The ***mixture of products with disjoint supports*** $\boldsymbol{\varrho}$ is the model $\mathcal{M}_\varrho := \operatorname{Mixt}(\overline{\mathcal{E}_{A_1}^1}, \dots, \overline{\mathcal{E}_{A_k}^1}) \subseteq \mathcal{M}_{n,k}$. For this kind of models, Lemmas 1 and 6 show:

Corollary 1. *Let $\varrho = \{A_1, \dots, A_k\}$ be a cubical partition of $\mathcal{X}$ with blocks $A_i = \mathcal{Y}_{i,1} \times \cdots \times \mathcal{Y}_{i,n}$ with $|\mathcal{Y}_{i,j}| \in \{1, q_i\}$ for all $j \in [n]$, for all $i \in [k]$. Then*

$$D_{\mathcal{M}_\varrho} = \max_{i \in [k]} \log(|A_i|/q_i)\ .$$

3.3 Unions of exponential families

Let $\mathcal{M}_{n,k}^* = \bigcup_{\varrho : |\varrho| = k} \mathcal{M}_\varrho \subseteq \mathcal{M}_{n,k}$ be the union of mixtures of products with disjoint supports ϱ, where ϱ runs over all cubical partitions of $\mathcal{X}$ with k blocks. The set $\mathcal{M}_{n,k}^*$ is not an exponential family, but a finite union of exponential families. Similarly, let $\mathcal{M}^*_{n,k,0} = \bigcup_{\varrho\colon|\varrho|=k}\mathcal{P}_\varrho$ be the union of all partition models $\mathcal{P}_\varrho$ of partitions ϱ with k cubical blocks.

Our motivation for studying unions of mixture models and unions of partition models comes from the following two results. For simplicity, we consider binary units; analogue results for non-binary units can be found in [15] and [12].

Theorem 6 ([17, Theorem 1]). *The binary model* $\mathrm{RBM}_{n,m}$ *contains any mixture of one arbitrary product distribution,* $m-k$ *product distributions with mutually disjoint supports, and* k *distributions with support on any edges of the* n*-cube, for any* $0 \leq k \leq m$*. In particular,* $\mathrm{RBM}_{n,m}$ *contains* $\mathcal{M}^*_{n,m+1}$*.*

Theorem 7 ([14, Theorem 17]). *Let* $L \in \mathbb{N}$*, let* k *be the largest integer for which* $L \geq 1 + 2^{(2^{k-1})}$*, and let* $K = 2^{k-1} + k \leq n$*. The binary deep belief network model with* L *layers of width* n *contains any partition model* $\mathcal{P}_\varrho$ *with partition* $\varrho = \{\{x\colon x_\lambda = y_\lambda\}\colon y_\lambda \in \{0,1\}^K\}$*, where* $\lambda \subseteq [n], |\lambda| = K$*.*

Unions of exponential families are more difficult to describe than exponential families, but the maximal rI-projection can be approximated as follows:

Theorem 8. *Let* $\mathcal{X} = \{0,1\}^n$*. If* $k \leq 2^{n-1}$*, then*

$$D_{\mathcal{M}^*_{n,k}} \leq \big(n - \lfloor\log_2(k)\rfloor - \frac{k}{2^{\lfloor\log_2(k)\rfloor}}\big)\log(2)\ .$$

If $k \leq 2^n$*, then*

$$D_{\mathcal{M}^*_{n,k,0}} \leq \big(n + 1 - \lfloor\log_2(k)\rfloor - \frac{k}{2^{\lfloor\log_2(k)\rfloor}}\big)\log(2)\ .$$

The first part was shown in [17, Theorem 2]. The second part can be proved with a direct adaptation of the same proof. Theorem 8, together with Theorems 6 and 7, proves the 'tighter bounds' in Theorems 1 and 2.

4 Discussion

When we plot the approximation error bounds of the model classes discussed here against the corresponding number of model parameters, we find that they all behave similarly; they all decay logarithmically on a large scale. This is the optimal maximal approximation error behaviour of exponential families (Theorem 4). The bounds for partition models, homogeneous independence models, and mixtures of products with disjoint homogeneous supports, are tight. The naïve Bayes model bound is tight for many choices of the N_i in the sense that it vanishes iff the model is a universal approximator, see [11]. The other bounds for the more complicated models are probably not tight. It is reasonable to expect that fixing the number of parameters, models with many hidden units fill the probability simplex more evenly than their counterparts with fewer or no hidden units (see, e.g., Lemma 3). For the discussed model classes, this paper does not give conclusive answers in that direction, since the only maximal divergence lower-bounds are for exponential families. It should be mentioned, however, that the mere existence of universal approximators within a given class of networks is not always obvious and sometimes false. For example, DBNs with too narrow hidden layers are never universal approximators, regardless of their parameter count.

Acknowledgement: J. R. is supported in part by the VW Foundation; G. M. by DARPA grant FA8650-11-1-7145.

References

1. N. Ay and A. Knauf. Maximizing multi-information. *Kybernetika*, 42:517–538, 2006.
2. N. Ay, G. Montúfar, and J. Rauh. Selection criteria for neuromanifolds of stochastic dynamics. In *Advances in Cognitive Neurodynamics (III)*. Springer, 2013.
3. G. Cybenko. Approximation by superpositions of a sigmoidal function. Technical report, Department of computer Science, Tufts University, Medford, MA, 1988.
4. K. Funahashi. Multilayer neural networks and Bayes decision theory. *Neural Networks*, 11(2):209 – 213, 1998.
5. K. Hornik, M. B. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.
6. J. Juríček. Maximization of information divergence from multinomial distributions. *Acta Universitatis Carolinae*, 52(1), 2011.
7. N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, 20(6):1631–1649, 2008.
8. N. Le Roux and Y. Bengio. Deep belief networks are compact universal approximators. *Neural Computation*, 22:2192–2207, 2010.
9. F. Matúš and N. Ay. On maximization of the information divergence from an exponential family. In *Proceedings of the WUPES'03*, pages 199–204, 2003.
10. F. Matúš. Maximization of information divergences from binary i.i.d. sequences. In *Proceedings IPMU*, pages 1303–1306, 2004.
11. G. Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *Kybernetika*, 49(1), 2013.
12. G. Montúfar. Universal approximation depth and errors of narrow belief networks with finite-valued units. Unpublished manuscript, 2013.
13. G. Montúfar and N. Ay. Refinements of universal approximation results for DBNs and RBMs. *Neural Computation*, 23(5):1306–1319, 2011.
14. G. Montúfar and J. Morton. Kernels and submodels of deep belief networks. 2012. Preprint available at `http://arxiv.org/abs/1211.0932`.
15. G. Montúfar and J. Morton. Discrete restricted Boltzmann machines. 2013. Preprint available at `http://arxiv.org/abs/1301.3529`.
16. G. Montúfar and J. Rauh. Scaling of model approximation errors and expected entropy distances. In *Proceedings of the WUPES'12*, pages 137–148, 2012.
17. G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In *Advances in NIPS 24*, pages 415–423, 2011.
18. J. Rauh. Finding the maximizers of the information divergence from an exponential family. *IEEE Transactions on Information Theory*, 57(6):3236–3247, 2011.
19. J. Rauh. Optimally approximating exponential families. *Kybernetika*, 2013. accepted. Preprint available at `http://arxiv.org/abs/1111.0483`.
20. I. Sutskever and G. E. Hinton. Deep narrow sigmoid belief networks are universal approximators. *Neural Computation*, 20:2629–2636, 2008.