The expressive power of mixture models and Restricted Boltzmann Machines

Johannes Rauh
joint work with Guido Montúfar and Nihat Ay

Max Planck Institute
Mathematics in the Sciences

Workshop on Graphical Models
April 2012 (Fields Institute, Toronto)

Outline

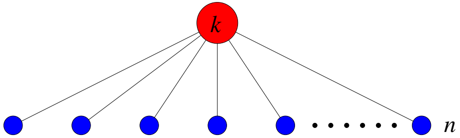
- Mixture Models and RBMs
- Submodels of RBMs
- The number of modes

Outline

- Mixture Models and RBMs
- Submodels of RBMs
- The number of modes

Mixture models

Consider n binary random variables. We want to study product distributions and their mixtures.



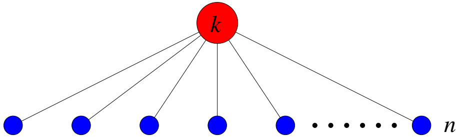
Definition

The kth *mixture model* $M_{n,k}$ consists of all convex combinations of k product distributions.

$$p(x_1, \ldots, x_n) = \sum_{i=1}^{k} \lambda_i q_{i,1}(x_1) q_{i,2}(x_2) \ldots q_{i,n}(x_n).$$

Mixture models

Consider n binary random variables. We want to study product distributions and their mixtures.



Definition

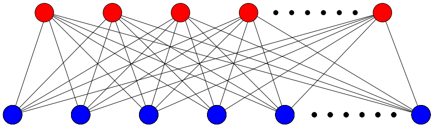
The kth *mixture model* $M_{n,k}$ consists of all convex combinations of k product distributions.

$M_{n,k}$ is *(the closure of)* a graphical model with one hidden node of size k.

With this definition, the first mixture model is the *independence model*.

Restricted Boltzmann Machines (RBMs)

Consider n binary random variables.



Definition

The *Restricted Boltzmann Machine* $\mathrm{RBM}_{n,m}$ is *(the closure of)* the graphical model of the complete bipartite graph $K_{n,m}$, where the group of m nodes is hidden.

Relations between mixture models and RBMs

- $\mathrm{RBM}_{n,m}$ is a submodel of $M_{n,2^m}$:

  $$p(v) = \sum_h p(v,h) = \sum_h p(v|h)p(h),$$

  where the conditionals $p(v|h)$ are product distributions.

  The *mixture components* $p(h)$ belong to $\mathrm{RBM}_{m,n}$.

- For $m = 1$, equality holds: $\mathrm{RBM}_{n,1} = M_{n,2}$.

- $\mathrm{RBM}_{n,m}$ equals the mth *Hadamard power* of $M_{n,2}$:
  - Hadamard product of functions = point-wise product
  - for probability distributions: renormalize afterwards

    (Observation due to Cueto, Morton and Sturmfels 2009)

The expressive power

We want to describe, *as precisely as possible,* which probability distributions a model does or does not contain.

- Both RBMs and mixture models are *semi-algebraic sets*: They have an implicit description in terms of *polynomial equations and inequalities*.

Question

How does this semi-algebraic description look like?

This description would allow to easily check whether a given distribution belongs to the model.

...but it appears to be too difficult to compute.

The expressive power

Since a complete description seems out of reach, we can ask other questions:

Easier problems

- What is the dimension of the model?
- Find large subsets of the model that are easy to describe.
- Find large sets of probability distributions that are not contained in the model.

The dimension of a model

In many cases, the dimension of a parametrically defined semi-algebraic set is the *expected dimension*, i.e. the number of parameters (or the dimension of the ambient space).

- The dimension of binary mixture models was recently computed:

Theorem (Catalisano, Geramita and Gimigliano 2011)
The dimension of $M_{n,k}$ equals the expected dimension $\min\{nk + k - 1, 2^n - 1\}$, unless $n = 4$ and $m = 3$.

- For RBMs, the dimension is as expected in all known cases:

Theorem (Cueto, Morton and Sturmfels 2009)
The dimension of $\mathrm{RBM}_{n,m}$ equals the expected dimension $\min\{nm + n + m, 2^n - 1\}$ for $k \leq 2^{n-\lceil \log_2(n+1) \rceil}$ and for $k \geq 2^{n-\lfloor \log_2(n+1) \rfloor}$.

The role of dimension

It is wellknown that the dimension alone is not sufficient to decide, whether a model is *full*, i.e. contains all probability distributions:

- Zwiernik and Smith computed all inequalities of $M_{3,2}$.
  Montúfar proved that $M_{n,k}$ is full if and only if $k \geq 2^{n-1}$.

  For $n = 3$:

k	1	2	3	4
dim	3	7	7	7
full?	no	no	no	yes

- For RBMs with $n = 3$:

m	0	1	2	3
dim	3	7	7	7
full?	no	no	no	yes

The model and its complement

The rest of the talk is about two projects that attack the following two problems:

Problem 1

Find large subsets of the model that are easy to describe.

We find “large” subsets of $\mathrm{RBM}_{n,m}$. These subsets are related to mixtures of product distributions on disjoint supports and allow to estimate the maximal approximation error.

Problem 2

Find large sets of probability distributions that are not contained in the model.

We find “large” sets outside of $M_{n,k}$. Interestingly, these sets touch the uniform distribution, showing that the uniform distribution need not be an interior point of $M_{n,k}$, even if $M_{n,k}$ has the full dimension.

Outline

- Mixture Models and RBMs
- Submodels of RBMs
- The number of modes

Cubical sets

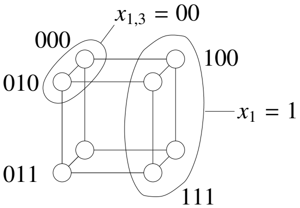
Idea:

- The RBM consists of mixtures of product distributions.
- Mixtures are difficult to describe...

...unless they have disjoint supports

Definition

A set $\mathcal{Y} \subseteq \{0,1\}^n$ is *cubical*, if it corresponds to a face of the n-dimensional hypercube.



Cubical sets are *cylinder sets*, i.e. they are characterized by "sub-configurations."

Cubical sets

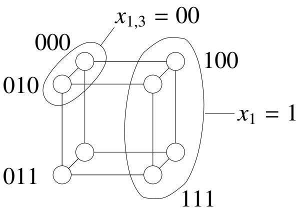
Idea:

- The RBM consists of mixtures of product distributions.
- Mixtures are difficult to describe...

...unless they have disjoint supports

Definition

A set $\mathcal{Y} \subseteq \{0,1\}^n$ is *cubical*, if it corresponds to a face of the n-dimensional hypercube.



Cubical sets are the *support sets* of product distributions.

Mixtures on disjoint supports

Theorem

$\mathrm{RBM}_{n,m}$ *contains any mixture of one arbitrary product distribution and* m *product distributions with pairwise disjoint cubical supports.*

Corollary

If $m \geq 2^{n-1} - 1$, *then* $\mathrm{RBM}_{n,m}$ *is full.*

Proof of the Corollary.

- Any distribution on an edge is a product distribution.
- The n-dimensional hypercube is covered by 2^{n-1} disjoint edges
- Hence any distribution is a mixture of 2^{n-1} product distributions supported on disjoint edges. □

The approximation error

Now we can find upper bounds for the *approximation error*:

Theorem

Let $m \leq 2^{n-1} - 1$. Then the Kullback-Leibler divergence from any distribution on $\{0,1\}^n$ to $\mathrm{RBM}_{n,m}$ is upper bounded by

$$\max_p D(p \| \mathrm{RBM}_{n,m}) \leq n - \lfloor \log(m+1) \rfloor - \frac{m+1}{2^{\lfloor \log(m+1) \rfloor}}$$

The bound gives an idea about the value of additional hidden nodes.

Idea of the proof:

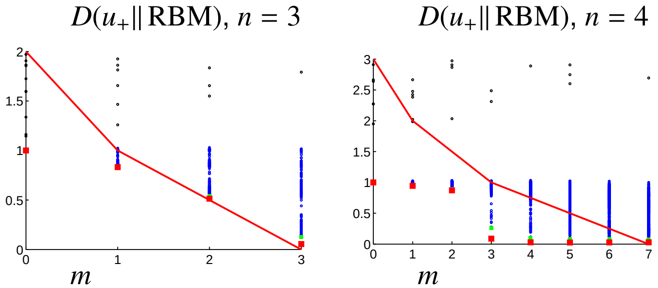
- Approximate a distribution p by a mixture of product distributions on disjoint supports.
- Where p puts more mass, the approximation must be better (choose smaller cubical sets).

Examples

Theorem

Let $m \leq 2^{n-1} - 1$. Then the Kullback-Leibler divergence from any distribution on $\{0,1\}^n$ to $\mathrm{RBM}_{n,m}$ is upper bounded by

$$\max_p D(p \| \mathrm{RBM}_{n,m}) \leq n - \lfloor \log(m+1) \rfloor - \frac{m+1}{2^{\lfloor \log(m+1) \rfloor}}$$



Outline

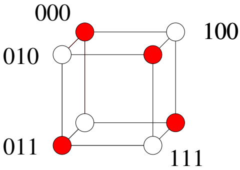
- Mixture Models and RBMs
- Submodels of RBMs
- The number of modes

The set Z_+

Question

How to prove that $M_{n,k}$ is not full if $k < 2^{n-1}$?

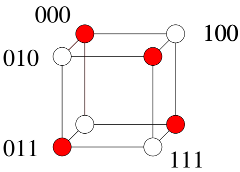
Denote $Z_\pm$ the elements of $\{0,1\}^n$ with *even/odd parity*.



Lemma

If $k < 2^{n-1}$, then $M_{n,k}$ does not contain the uniform distribution u_+ on Z_+.

The set Z_+



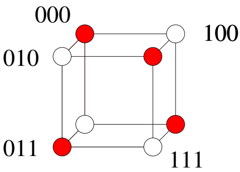
Lemma

If $k < 2^{n-1}$, then $M_{n,k}$ does not contain the uniform distribution u_+ on Z_+.

Proof.

- If $u_+ = \sum_{i=1}^{N} \lambda_i p_i$, with $\lambda_i > 0$, then $Z_+ = \cup_i \operatorname{supp}(p_i)$.
- If the p_i are product measures, then $\operatorname{supp}(p_i)$ is cubical.
- Only the one-element subsets of Z_+ are cubical. $\square$

The set Z_+



Note: $M_{n,k}$ is full if and only if $M_{n,k}$ contains u_+.

Conjecture

The same is true for RBMs: $\mathrm{RBM}_{n,m}$ is full if and only if $\mathrm{RBM}_{n,m}$ contains u_+.

(true for $n = 3$)

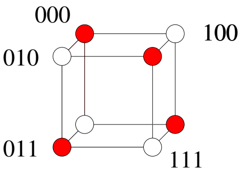
The number of modes

Idea: Find a neighbourhood of u_+ which is not contained in $M_{n,k}$.

Definition

A *mode* of a distribution p is a strict local maximum of p, where "local" refers to the neighbourhood structure on the cube.

- A single product distribution has (at most) one mode.
- u_+ has 2^{n-1} modes.
- If p has 2^{n-1} modes, then the set of modes equals Z_+ or Z_-.



The number of modes

Idea: Find a neighbourhood of u_+ which is not contained in $M_{n,k}$.

Definition

A *mode* of a distribution p is a strict local maximum of p, where "local" refers to the neighbourhood structure on the cube.

- A single product distribution has (at most) one mode.
- u_+ has 2^{n-1} modes.
- If p has 2^{n-1} modes, then the set of modes equals Z_+ or Z_-.

Question

How many modes can a mixture of k product distributions have?

The number of modes

Let $\alpha(n,k)$ denote the maximum number of modes that $p \in M_{n,k}$ may have.

Properties:

- $2^{n-1} \geq \alpha(n,k) \geq \min\{k, 2^{n-1}\}$
- $\alpha(n,1) = 1$
- $\alpha(3,2) = 2$
- $\alpha(3,3) = 3$

Corollary

$M_{3,3}$ is not full.

Distributions with four modes

Result

$\alpha(3,3) = 3$, and hence $M_{3,3}$ is not full.

Note that there are distributions with four modes arbitrarily close to the uniform distribution.

Corollary

The uniform distribution is not an interior point of $M_{3,3}$, even though $M_{3,3}$ is full-dimensional.
In particular, the uniform distribution is a singularity of $M_{3,3}$.

The set of distributions with four modes is a union of two polyhedral sets, containing distributions with four modes on $Z_\pm$. Both polyhedral parts touch the uniform distributions.

The number of modes

Let $\alpha(n,k)$ denote the maximum number of nodes that $p \in M_{n,k}$ may have.

Properties:

- $2^{n-1} \geq \alpha(n,k) \geq \min\{k, 2^{n-1}\}$
- $\alpha(n,1) = 1$
- $\alpha(3,2) = 2$
- $\alpha(3,3) = 3$

The number of modes

Let $\alpha(n,k)$ denote the maximum number of nodes that $p \in M_{n,k}$ may have.

Properties:

- $2^{n-1} \geq \alpha(n,k) \geq \min\{k, 2^{n-1}\}$
- $\alpha(n,1) = 1$
- $\alpha(3,2) = 2$
- $\alpha(3,3) = 3$
- $\alpha(4,2) = 3$

This tells us:

It is not sufficient to consider the number of modes:
For example, $\alpha(4,7) = 8$, but $M_{4,7}$ is not full.

Summary

- Mixture models and RBM are important statistical models with many open problems.
- Mixtures of product distributions with disjoint supports can help to understand RBMs.
- Distributions with the maximal number of modes are difficult to approximate with mixture models and RBMs.

Open Problems:

- Compute $\alpha(n, k)$ for $n \geq 4$, $k \geq 2$.
- Is the uniform distribution always a singularity of $M_{n,k}$ (unless the model is full)?
- What about $\mathrm{RBM}_{n,m}$?

further reading

G. Montúfar, J. Rauh, N. Ay
Expressive Power and Approximation Errors of RBMs
NIPS 2011

G. Montúfar
Mixture Decompositions using a Decomposition of the Sample Space
arXiv 1008.0204

A. Cueto, J. Morton, B. Sturmfels
Geometry of the Restricted Boltzmann Machine
Algebraic Methods in Statistics and Probability II

P. Zwiernik, J. Smith
Implicit inequality constraints in a binary tree model
Electronic Journal of Statistics