Some RBM Geometry

Jason Morton

Penn State

June 13, 2018
TSIMF

Overview

- A bit of work on "Mathematical Foundations of Deep Learning" beginning in 2008 (+2010 DARPA program), series of results on geometry and representational power of RBM.
- Hiring postdoc (3 so far, now at UCLA, Skolkovo, U.S. Naval Academy)
- First a word about tensor networks

Tensor networks

Tensor networks specialize to quantum and statistical versions

- Neural networks and graphical models, statistical physics models share origins
- Formalize with parameterizations (MPS paper, MASS course)
- Formalize with monoidal category theory (2014 lecture series at IMA)
- Software?

Approximate Dictionary [Critch M- 2012], Benasque

Tensor Networks in Physics	Graphical Models in Stats/ML
MPS	HMM
TTN	GMM
PEPS	CRF/MRF
MERA	?DBM?
DMRG	Forward-backward algorithm

In Algebraic Statistics we have been studying the right-hand column

- often determining the ideal / variety / manifold / orbit structure and characteristics (e.g. identifiability) of the parameterization
- generally work in complex projective space
  - so pure states are more natural than probabilities
- related optimization, contraction, approximation problems

Approximate Dictionary

Tensor Networks in Physics	Graphical Models in Stats/ML
MPS	**HMM**
TTN	GMM
PEPS	CRF/MRF
MERA	?DBM?
DMRG	Forward-backward algorithm

In Algebraic Statistics we have been studying the right-hand column

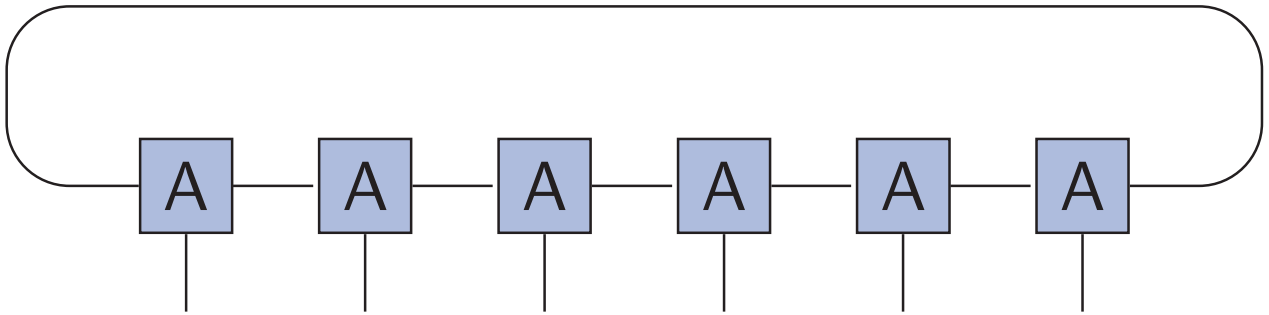
- often determining the ideal / variety / manifold / orbit structure and characteristics (e.g. identifiability) of the parameterization
- generally work in complex projective space
  - so pure states are more natural than probabilities
- related optimization, contraction, approximation problems

Matrix product states as tensor networks

Fix parameter matrices $A_1, \ldots, A_d$.

$$\Psi = \sum_{i_1,\ldots,i_n} \operatorname{tr}(A_{i_1} \cdots A_{i_n}) |i_1 i_2 \cdots i_n\rangle$$

is a translation-invariant matrix product state with periodic boundary conditions



Which states are matrix product states?

Matrix product states as algebraic varieties

Fix parameter matrices $A_1, \ldots, A_d$.

$$\Psi = \sum_{i_1,\ldots,i_n} \mathrm{tr}(A_{i_1} \cdots A_{i_n}) | i_1 i_2 \cdots i_n \rangle$$

- Which states are matrix product states? More precisely,
- What is the variety of MPS (closure of the image of this parameterization), and the ideal I of polynomial relations that hold among the coefficients

$$\Psi_{i_1,\ldots i_n} = \mathrm{tr}(A_{i_1} \cdots A_{i_n})?$$

- This implicitization is the space of quantum states representable as matrix product states.

Binary Matrix Product States

For $N \leq 3$ states, there are no relations; for $N = 4$, a hypersurface

Theorem (Critch-M- 2012)

A four qubit state Ψ is a binary periodic translation invariant MPS if and only if if the following irreducible polynomial vanishes:*

$$\psi_{1010}^2\psi_{1100}^4 - 2\psi_{1100}^6 - 8\psi_{1000}\psi_{1010}\psi_{1100}^3\psi_{1110} + 12\psi_{1000}\psi_{1100}^4\psi_{1110} - 4\psi_{1000}^2\psi_{1010}^2\psi_{1110}^2 + 2\psi_{0000}\psi_{1010}^3\psi_{1110}^2 + 16\psi_{1000}^2\psi_{1010}\psi_{1100}\psi_{1110}^2 - 4\psi_{0000}\psi_{1010}^2\psi_{1100}\psi_{1110}^2 - 16\psi_{1000}^2\psi_{1100}^2\psi_{1110}^2 + 4\psi_{0000}\psi_{1010}\psi_{1100}^2\psi_{1110}^2 - 4\psi_{0000}\psi_{1100}^3\psi_{1110}^2 - 4\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1110}^3 + 8\psi_{0000}\psi_{1000}\psi_{1100}\psi_{1110}^3 - \psi_{0000}^2\psi_{1110}^4 + 2\psi_{1000}^2\psi_{1010}^3\psi_{1111} - \psi_{0000}\psi_{1010}^4\psi_{1111} - 4\psi_{1000}^2\psi_{1010}^2\psi_{1100}\psi_{1111} + 4\psi_{1000}^2\psi_{1010}\psi_{1100}^2\psi_{1111} + 2\psi_{0000}\psi_{1010}^2\psi_{1100}^2\psi_{1111} - 4\psi_{1000}^2\psi_{1100}^3\psi_{1111} + \psi_{0000}\psi_{1100}^4\psi_{1111} - 4\psi_{1000}^3\psi_{1010}\psi_{1110}\psi_{1111} + 4\psi_{0000}\psi_{1000}\psi_{1010}^2\psi_{1110}\psi_{1111} + 8\psi_{1000}^3\psi_{1100}\psi_{1110}\psi_{1111} - 8\psi_{0000}\psi_{1000}\psi_{1010}\psi_{1100}\psi_{1110}\psi_{1111} - 2\psi_{0000}\psi_{1000}^2\psi_{1110}^2\psi_{1111} + 2\psi_{0000}^2\psi_{1010}\psi_{1110}^2\psi_{1111} - \psi_{1000}^4\psi_{1111}^2 + 2\psi_{0000}\psi_{1000}^2\psi_{1010}\psi_{1111}^2 - \psi_{0000}^2\psi_{1010}^2\psi_{1111}^2.$$

Of course it gets worse quickly

Theorem

Any homogeneous minimal generating set for the ideal of $\overline{cmps(2,2,5)}$ *must contain exactly 3 quartic and 27 sextic polynomials, possibly some higher degree polynomials, but none of degree* 1, 2, 3, *or* 5.

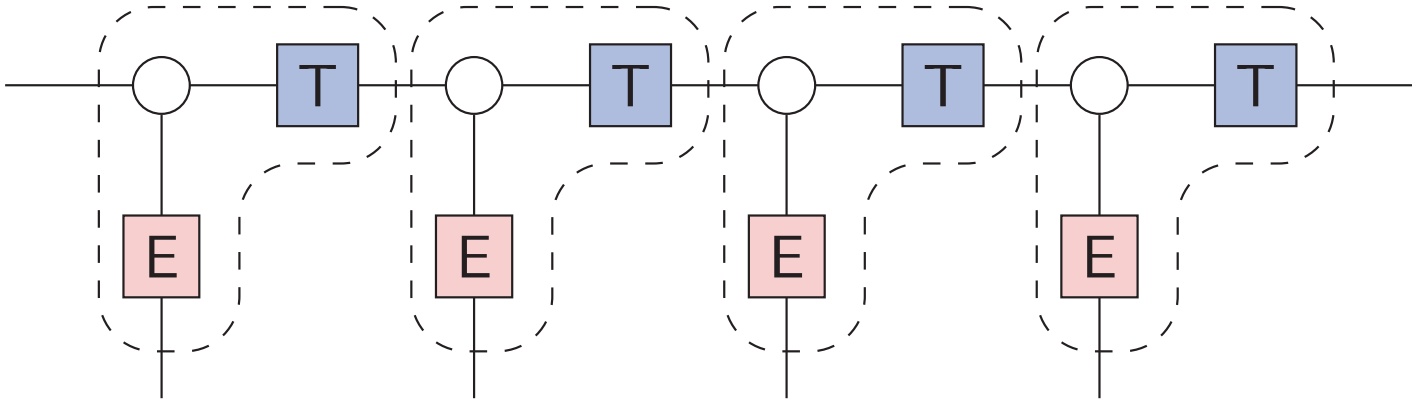
Theorem

A three-qubit state Ψ is a limit of $N = 3$ binary translation invariant MPS with open boundary conditions if and only if the following 22-term quartic polynomial vanishes:

$$
\begin{aligned}
&\psi_{011}^2\psi_{100}^2 - \psi_{001}\psi_{011}\psi_{100}\psi_{101} - \psi_{010}\psi_{011}\psi_{100}\psi_{101} \\
&+ \psi_{000}\psi_{011}\psi_{101}^2 + \psi_{001}\psi_{010}\psi_{011}\psi_{110} - \psi_{000}\psi_{011}^2\psi_{110} \\
&- \psi_{010}\psi_{011}\psi_{100}\psi_{110} + \psi_{001}\psi_{010}\psi_{101}\psi_{110} \\
&+ \psi_{001}\psi_{100}\psi_{101}\psi_{110} - \psi_{000}\psi_{101}^2\psi_{110} - \psi_{001}^2\psi_{110}^2 \\
&+ \psi_{000}\psi_{011}\psi_{110}^2 - \psi_{001}\psi_{010}^2\psi_{111} + \psi_{000}\psi_{010}\psi_{011}\psi_{111} \\
&+ \psi_{001}^2\psi_{100}\psi_{111} + \psi_{010}^2\psi_{100}\psi_{111} - \psi_{000}\psi_{011}\psi_{100}\psi_{111} \\
&- \psi_{001}\psi_{100}^2\psi_{111} - \psi_{000}\psi_{001}\psi_{101}\psi_{111} + \psi_{000}\psi_{100}\psi_{101}\psi_{111} \\
&+ \psi_{000}\psi_{001}\psi_{110}\psi_{111} - \psi_{000}\psi_{010}\psi_{110}\psi_{111}
\end{aligned}
$$

Previously seen in the context of the HMM [Pachter Sturmfels 2004].

MPS as quantum hidden Markov models



- T and E are complex matrices with rows summing to $z \in \mathbb{C}$, so the parameter space is a $\mathbb{C}^5$.
- The image of this parameteriztion is dense in the five-dimensional periodic boundary MPS $\overline{PB}(2, 2, N)$ for all N.
- Another way to formalize the analogy is in terms of monoidal categories.

Restricted Boltzmann Machine Papers

- Geometry of the Restricted Boltzmann Machine [0908.4425 / Algebraic Methods in Statistics and Probability 2010] with Cueto and Sturmfels.
- Discrete Restricted Boltzmann Machines [1301.3529 / ICLR 2013 / JMLR 2015] with Montufar.
- When does a mixture of products contain a product of mixtures? [1206.0387 / SIAM J. Discrete Math 2015] with Montufar.
- Dimension of Marginals of Kronecker Product Models [1511.03570 / SIAM J. Appl. Algebraic Geometry 2017] with Montufar.

Expected dimension

- A model having the expected dimension (i.e., not being defective) is
  - a necessary condition for the model parameters to be identifiable and
  - for much of traditional asymptotic theory such as asymptotic normality of estimators, convergence rates, and so on.
- It also means that no parameters are wasted: every degree of freedom entering as a parameter adds a dimension to the space of representable probability distributions, until that space is full-dimensional.

Geometry of the Restricted Boltzmann Machine

- Dimension for naïve Bayes with binary visible states was open until recently
- Coding theory proof with a gap near full dimensionality
- Conjecture: RBM always has the expected dimension (no. of parameters or ambient dimension)
- Cueto-Yu 2010, 22-2222 is a hypersurface, Newton Polytope has 17 million vertices
- 3-222 vs 22-222, Montufar-Siegal 2018
- That the RBM has expected dimension is a corollary of a bigger theorem: allowed two exponential families (one hidden, one visible), the RBM is the case where they are both independence models.
  - Works by generalizing hyperplane arrangement logic to an arrangement of normal fans of polytopes.

Kronecker product models

A Kronecker product model is a marginal exponential family model $\mathcal{M}_F$

$$p_\theta(x) = \sum_{y \in \mathcal{Y}} \frac{1}{Z(\theta)} \exp\left(\langle \theta, F(x,y) \rangle\right), \quad \text{for all } x \in \mathcal{X}, \quad \text{for all } \theta \in \mathbb{R}^d,$$

with factorizing sufficient statistics matrix $F(x,y) = A(x) \otimes B(y)$.

$$F = A \otimes B := \begin{pmatrix} a_{1,1}B & a_{1,2}B & \cdots & a_{1,m}B \\ a_{2,1}B & \ddots & & \\ \vdots & & & \\ a_{n,1}B & & & a_{n,m}B \end{pmatrix}$$

Product of "visible" exponential family $\mathcal{E}_A$ on $\mathcal{X}$ and "hidden" exponential family $\mathcal{E}_B$ on $\mathcal{Y}$.

Mixture of products $\mathcal{M}_{n,k}$

- Probability distributions on n binary variables $X_1, \ldots, X_n$
- Product distribution factors: $p(x) = p(x_1) \cdots p(x_n)$
- A k-mixture of products $\mathcal{M}_{n,k}$, a naïve Bayes model, is
  - a convex combination of k product distributions
  - Closure of the n-dim exponential family $p_B(x) = \frac{1}{Z(B)} \exp(B^\top x)$
  - k pockets each with n differently-colored biased coins

Product of mixtures of products: $\mathrm{RBM}_{n,m}$

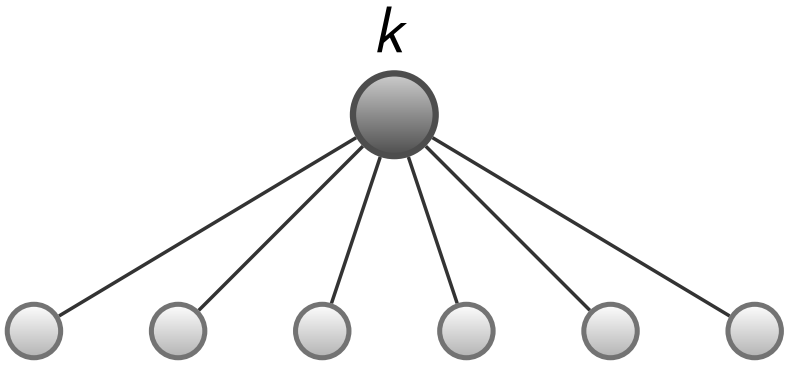
- Mixture of experts: Hadamard (pointwise) product of m different $\mathcal{M}_{n,2}$s.
- The RBM model with n visible and m hidden binary units is the set $\mathrm{RBM}_{n,m}$ of distributions on $\{0,1\}^n$ which are limits of

$$p(x) = \frac{1}{Z(W,B,C)} \sum_{h \in \{0,1\}^m} \exp(h^\top W x + B^\top v + C^\top h),$$

  - visible units $v \in \{0,1\}^n$, hidden units $h \in \{0,1\}^m$
  - $W \in \mathbb{R}^{m \times n}$ is a matrix of interaction weights
  - $B \in \mathbb{R}^n$ is the visible bias weights
  - $C \in \mathbb{R}^m$ is the hidden bias weights
  - $Z = \sum_{v,h} \exp(h^\top W x + B^\top x + C^\top h)$ normalizes
- "Usually" has the expected dimension [Cueto M Sturmfels].

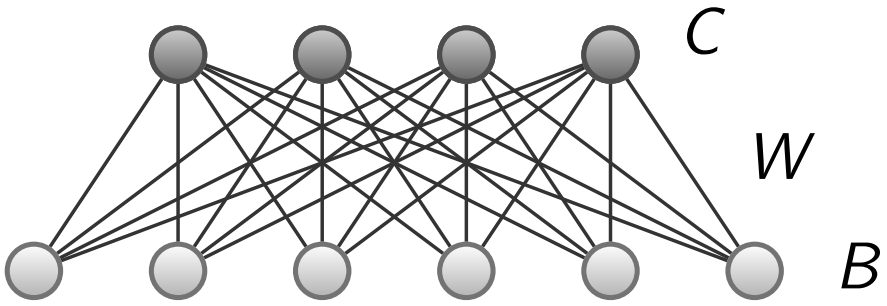
How are RBMs better?

Mixture of Products



$\mathcal{M}_{6,k}$

Product of Mixtures of Products (RBM)



$\mathrm{RBM}_{6,4}$

Mixture of products and product of mixtures as graphical models. Dark nodes represent hidden units, light nodes represent visible units.

Problem (1)

When does the mixture of product distributions $\mathcal{M}_{n,k}$ contain the product of mixtures of product distributions $\mathrm{RBM}_{n,m}$, and vice versa?

- Theorem: Fixing $\frac{m}{n} \in (0, \infty)$, the number of parameters of
  - the smallest mixture of products $\mathcal{M}_{n,k}$ containing $\mathrm{RBM}_{n,m}$
  - grows exponentially in the number of parameters of the RBM
- Because the RBM can represent distributions with many more Hamming-local maxima
- Comes from polyhedral approximations of the sets of probability distributions representable by each model.

Related problems

Problem

What sets of length-n binary vectors are

(2) *perfectly reconstructible by an RBM with m hidden units?*

(3) *the outputs of n linear threshold functions with m input bits?*

(4) *the modes or strong modes (Hamming-local maxima) of distributions represented by an RBM with m hidden units?*

- *Probability distributions with many strong modes (e.g. w/even parity support) are easy for RBMs but hard for naïve Bayes.*

Modes and hyperplanes

- Modes described by linear inequalities of the form $p(x) > p(x')$, defining polyhedral approximations of probability models.
- Closely related to binary classification problems and separation of vertex sets of hypercubes by hyperplane arrangements, leading to problems such as

Problem (5)

What is the smallest arrangement of hyperplanes, if one exists, that slices each edge of a hypercube a given number of times?

- These five problems are nearly equivalent.
- DRBM and Kronecker papers generalize these ideas to non-binary variables and allowing interactions within layers.
- Each hyperplane becomes a normal fan, n-cube becomes product of simplices

Inference functions of RBMs

- The inference function of a probability model $p_\theta(v, h)$ with parameter $\theta \in \Omega$ explains each visible value v by the most likely hidden value of h according to

$$\begin{aligned} \mathrm{up}_\theta : v &\mapsto \operatorname{argmax}_h p_\theta(h|v) \\ \mathrm{up}_\theta : \{0,1\}^n &\to \{0,1\}^m \\ \mathrm{up}_\theta : \mathbb{R}^n &\to \{0,1\}^m \end{aligned}$$

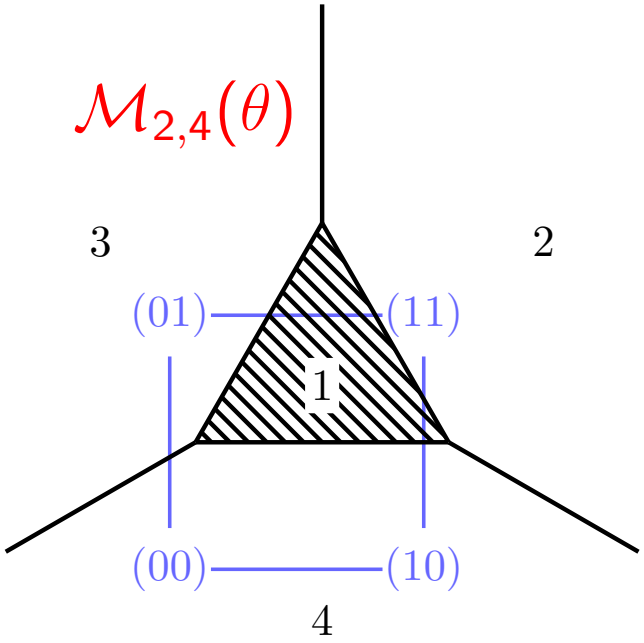
- Each of the m RBM hidden units linearly divides the input space $\mathbb{R}^n$ according to its preferred state given the input.

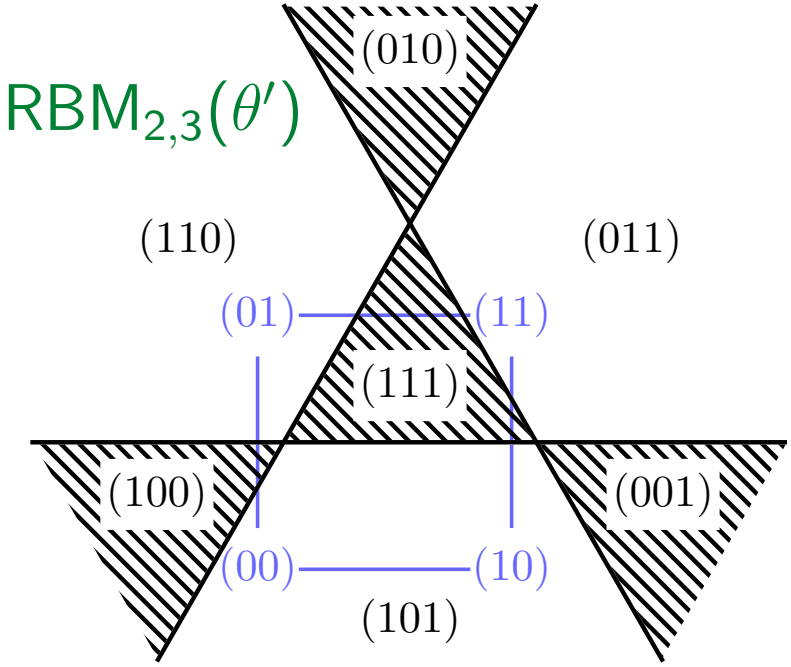
Inference regions and distributed representations

$$\begin{aligned} \mathrm{up}_\theta \colon \{0,1\}^n &\to \{0,1\}^m \\ \mathrm{up}_\theta \colon \mathbb{R}^n &\to \{0,1\}^m \end{aligned}$$

- The m hidden units partition input space into inference regions where different **joint** hidden states $h_1, \ldots, h_m$ are most likely.
- Such a distributed representation is speculated [Bengio 2009] to be a key to the model's efficacy.

Inference regions of mixture and RBM, for some θ





- Both $\mathcal{M}_{2,4}$ and $\mathrm{RBM}_{2,3}$ have 7 parameters
- Both can approximate any distribution on $\{0,1\}^2$.
- Define very different inference regions.

RBM combines linear threshold inference functions

Definition

A *linear threshold function* (LTF) with m (binary) inputs is a function

$$f\colon \{0,1\}^m \to \{-,+\}; \quad y \mapsto \operatorname{sgn}((\sum_{j\in[m]} w_j y_j) + b);$$

where $w \in \mathbb{R}^m$ is called *weight vector* and $b \in \mathbb{R}$ *bias*.

- Identify $-/+$ and $0/1$ vectors via $- \leftrightarrow 0$ and $+ \leftrightarrow 1$.
- Choosing parameters $W \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^n$, $C \in \mathbb{R}^m$, our model $\mathrm{RBM}_{n,m}$ defines the inference function

$$\mathrm{up}_{W,B,C}\colon\ \mathbb{R}^n \supset \{0,1\}^n \to \{0,1\}^m;\ v \mapsto \operatorname{argmax}_{h\in\{0,1\}^m} h^\top(Wv+C)\ .$$

- The log number of LTFs with m inputs is asymptotically of order m^2, the exact number is only known for $m \leq 9$.

Inference functions: linear threshold functions

- Partition given by intersection of an affine space and the normal fan of an m-cube (the orthants of $\mathbb{R}^m$).
  - preimages of the orthants of $\mathbb{R}^m$ by the affine map $\psi\colon \mathbb{R}^n \to \mathbb{R}^m; v \mapsto Wv + C$.
- Number of inference regions $\leq \sum_{i=0}^{\mathrm{rank}(W)} \binom{m}{i}$ $=$ number of orthants of $\mathbb{R}^m$ intersected by a generic d-dimensional affine subspace.
- When $\mathrm{rank}(W) < m$ (e.g. if $m > n$),
  - the image of the map ψ does not intersect all orthants of $\mathbb{R}^m$
  - there are 'empty' inference regions, i.e., states h which are not the explanation of any input vector v.

Inference functions of mixture model

The mixture model $\mathcal{M}_{n,k}$ has a simpler inference function

$$\mathrm{up}_{\lambda,B}: \quad v \mapsto \mathrm{argmax}_{i\in[k]}(B_i^\top v - \log(Z(B_i)) + \log(\lambda_i)), \qquad (1)$$

with mixture weights λ_i, parameters of each mixture component $B_i \in \mathbb{R}^n$ for $i \in [k]$, and $Z(B_i) = \sum_{v\in\{0,1\}^n} \exp(B_i^\top v)$.

- Input space $\mathbb{R}^n$ partitioned into at most k regions of linearity of the function $v \mapsto \max\{B_i^\top v - \log(Z(B_i)) + \log(\lambda_i) \colon i \in [k]\}$.
- Partition given by the intersection of an affine space and the normal fan of a $(k-1)$-simplex.

Comparing inference functions

- Fixing input space dimension n, the number of inference regions in $\mathbb{R}^n$ realizable by $\mathrm{RBM}_{n,m}$ is of order $\Theta(\binom{m}{\min\{n,m\}})$
- Exponential in the number of parameters of the model
- vs. number of inference regions realizable by mixture $\mathcal{M}_{n,k}$: linear in the number of parameters of the model.

So distributed representations can learn different explanations to a number of observations that is exponential in the number of model parameters.

Now let's look at the codes (sets of bitstrings) an RBM prefers.

Six 3-parameter properties of sets of binary vectors

Let **n**, **m** be non-negative integers and $\mathcal{C}$ be a subset of $\{0,1\}^n$.

- **LTC**: The set $\mathcal{C}$ is an (n, m)-**l**inear **t**hreshold **c**ode, i.e., the image of n linear threshold functions with m inputs.
- **HP**: There exists an arrangement $\mathcal{A}$ of n **h**yper**p**lanes in $\mathbb{R}^m$ such that the vertices of the m-dimensional unit cube intersect exactly the $\mathcal{C}$-cells of $\mathcal{A}$.
- **ZP**: There is an affine image of the vertices of an m-cube (m-**z***onoset*) in $\mathbb{R}^n$ which intersects exactly the $\mathcal{C}$-orthants of $\mathbb{R}^n$.
- **SM**: An RBM with n visible and m hidden nodes can represent a distribution with set of **s**trong **m**odes $\mathcal{C}$.
- **PR**: The set $\mathcal{C}$ is the set of **p**erfectly **r**econstructible inputs of an RBM with n visible and m hidden nodes.
- **SP**: An RBM with n visible and m hidden nodes can represent a distribution which is **s**trictly **p**ositive on $\mathcal{C}$ and zero elsewhere.

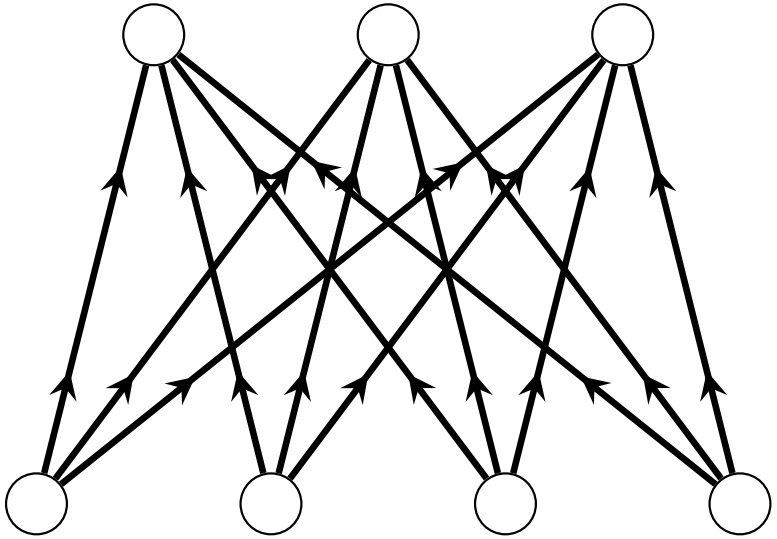
Perfect reconstructibility

- Similarly to the up_θ inference function, a model $p_\theta(v, h)$ defines a down_θ inference function,
- down_θ outputs the most likely visible state $\text{argmax}_v\, p_\theta(v|h)$ given a hidden state h.

Definition

Given a probability model $p_\theta(v, h)$ on $v \in \mathcal{X}$ and $h \in \mathcal{Y}$, a collection of states $\mathcal{C} \subseteq \mathcal{X}$ is *perfectly reconstructible* if there is a choice of the parameter θ for which $\text{down}_\theta(\text{up}_\theta(v)) = v$ for all $v \in \mathcal{C}$.

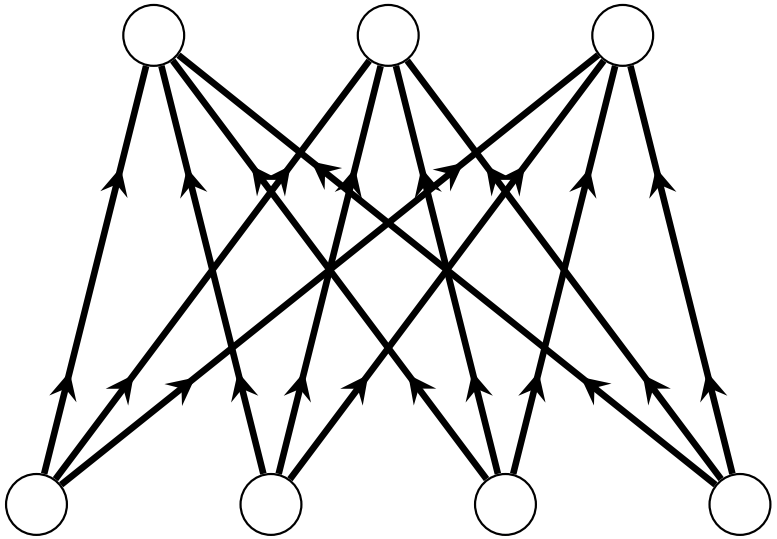
Is $\mathcal{C}$ reconstructible for some θ?



$$\mathcal{C} \quad \left\{ \begin{array}{cccc} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{array} \right.$$

Is $\mathcal{C}$ reconstructible for some θ?

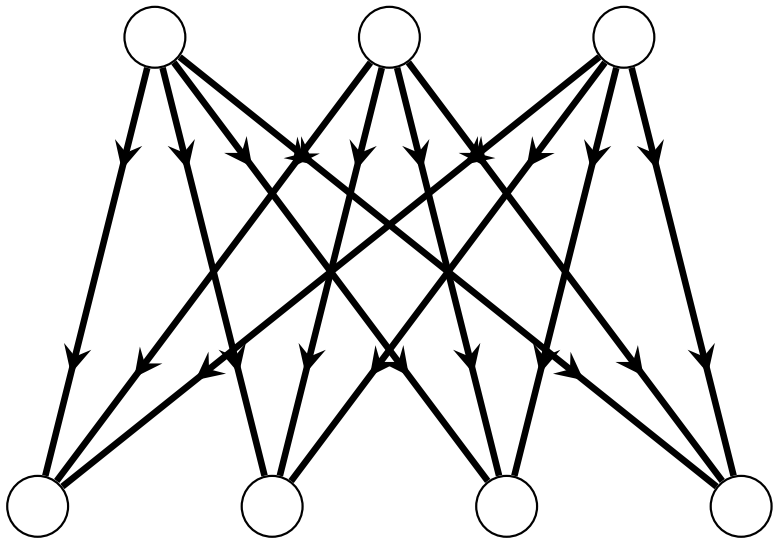
$$\mathrm{up}_\theta(\mathcal{C})\left\{\begin{array}{ccc} 1 & 0 & 1 \\ 0 & 0 & 1 \end{array}\right.$$



$$\mathcal{C}\quad\left\{\begin{array}{cccc} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{array}\right.$$

Is $\mathcal{C}$ reconstructible for some θ?

$$\mathrm{up}_\theta(\mathcal{C})\left\{\begin{matrix} 1 & 0 & 1 \\ 0 & 0 & 1 \end{matrix}\right.$$



$$\mathrm{down}_\theta(\mathrm{up}_\theta(\mathcal{C}))\left\{\begin{matrix} 0 & 0 & 1 & 1 \\ 1 & 1 \neq 0 & 0 & 1 \end{matrix}\right.$$

Perfect reconstructibility

- Ability to reconstruct input vectors
  - used to evaluate the performance of RBMs in practice
  - intuitive, can be tested more cheaply than probability distributions.
- Taking this seriously leads to autoencoder-like training algorithms, where we minimize reconstruction error.

Which subsets of $\{0,1\}^n$ can be reconstructed?

- Write the joint distribution on hidden and visible states

$$\{p_\theta(v,h)\}_{v,h}$$

  as a matrix with rows labeled by h and columns by v.
- Then $\mathcal{C}$ is perfectly reconstructible iff for some θ, we have $p_\theta(v, \mathrm{up}_\theta(v))$ is the unique maximal entry in the $\mathrm{up}_\theta(v)$-row (and in the v-column) for all $v \in \mathcal{C}$.

Represent distributions with many modes?

Let $d_H(x, y)$ be the Hamming distance between binary strings x and y: the number of bits that must be flipped to turn x into y.

Definition

A binary vector x is a *mode* of a distribution p if

$$p(x) > p(y) \text{ for all } y \in \mathcal{X} \text{ with } d_H(y, x) = 1,$$

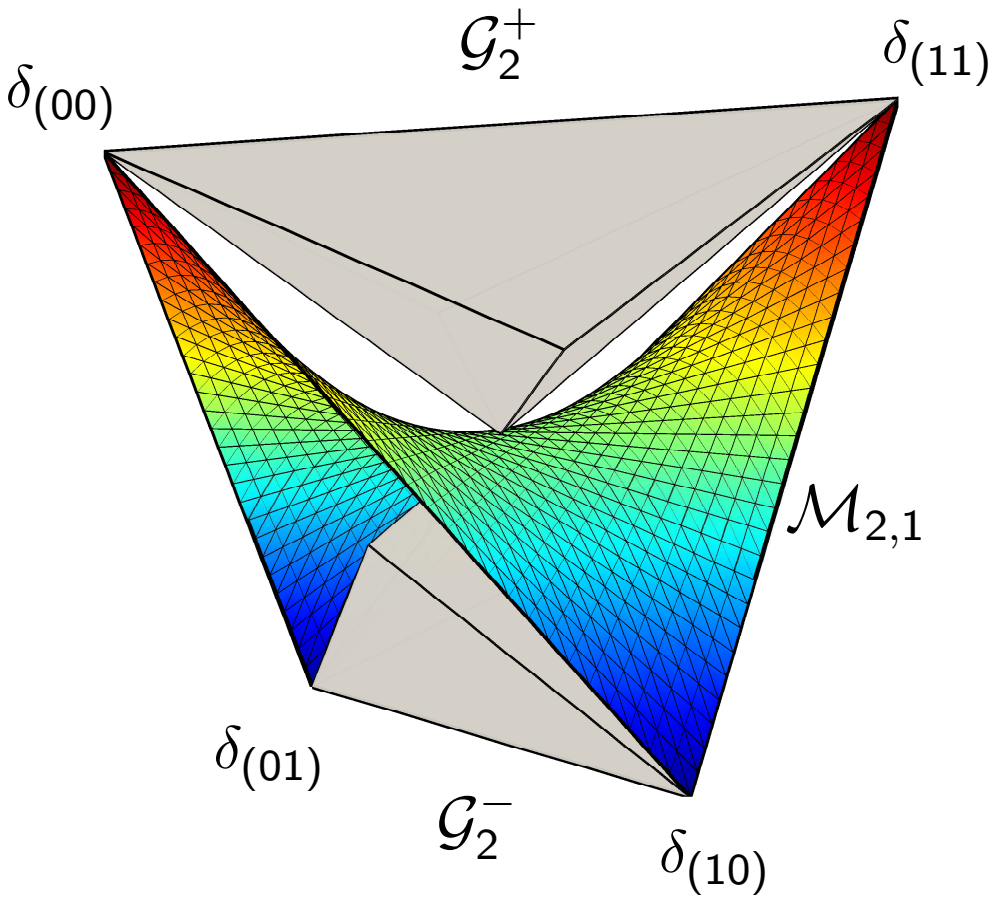
and a *strong mode* if $p(x) > \sum_{y \in \mathcal{X}: d_H(y,x)=1} p(y)$.

- The modes of a distribution are the Hamming-locally most likely events in the space of possible events.
- Modes are closely related to the support sets and boundaries of statistical models, which have been studied especially for hierarchical and graphical models without hidden variables.
- A "complex" distribution has many modes.

Polyhedral approximation with (strong) modes

- Consider the set of distributions which have modes $\mathcal{G}_{\mathcal{C}}$ or strong modes $\mathcal{H}_{\mathcal{C}}$ **exactly** at the bitstrings in a code $\mathcal{C}$ (so codewords Hamming distance 2 apart).
- The closures ($\overline{\mathcal{G}_{\mathcal{C}}}$ and $\overline{\mathcal{H}_{\mathcal{C}}}$ resp) are convex "mode polytopes" inscribed in the probability simplex
- The sets of modes not realizable by a probability model give a full-dimensional polyhedral approximation of the model's complement.
- Test cases: binary strings with an even or odd number of ones.
  - These are the maximal subsets of $\{0,1\}^n$, with cardinality 2^{n-1} and minimum distance two.

Mode polytopes on $\{0,1\}^2$



The polytopes at the top and bottom contain the distributions with two modes on even or odd parity bitstrings respectively.

Modes of mixtures of products

Theorem

- *The sets $\mathcal{C}$ of strong modes of distributions in $\mathcal{M}_{n,k}$ are exactly the sets of strings in $\mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ of minimum Hamming distance at least two and cardinality at most k.*
- *Furthermore, if $p \in \mathcal{M}_{n,k}$ has strong modes $\mathcal{C}$, then every $c \in \mathcal{C}$ is the mode of one mixture component of p.*

Fun fact: although the mixture $\mathcal{M}_{3,3}$ is full dimensional in Δ_{2^3-1}, its complement contains points arbitrarily close to the uniform distribution.

Modes of RBMs

Problem

What is the smallest $m \in \mathbb{N}$ for which $\mathrm{RBM}_{n,m}$ contains a distribution with l (strong) modes?

- In particular, what is the smallest m for which the model $\mathrm{RBM}_{n,m}$ can represent the parity function?
- Characterizing the sets of modes realizable by RBMs is a more complex problem for mixtures of product distributions
- Useful to describe in terms of point configurations called zonosets, hyperplane arrangements, or linear threshold functions.

Zonosets

- Zonotopes are equivalently images of n-cubes under affine projection, or Minkowski sums of line segments, and encode hyperplane arrangements.
- Zonosets remember the generating points (interior, multiplicity).
- Parameters W, B define the projection.

Definition

Let $m \geq 0$, $n > 0$, $W_i \in \mathbb{R}^n$ for all $i \in [m]$, and $B \in \mathbb{R}^n$. The multiset $\mathcal{Z} = \{\sum_{i \in I} W_i + B\}_{I \subseteq [m]}$ is called an *m-zonoset*.

- The convex hull of a zonoset is a zonotope.

Theorem: Connecting properties of codes

Let $\mathcal{C} \subset \{0,1\}^n$ be a binary code of minimum Hamming distance at least two.

- If the model $\mathrm{RBM}_{n,m}$
  - contains a distribution with strong modes $\mathcal{C}$
  - or $\mathcal{C}$ has cardinality 2^m and is perfectly reconstructible by $\mathrm{RBM}_{n,m}$,
  - then there is an m-zonoset with a point in each $\mathcal{C}$-orthant of $\mathbb{R}^n$.
- If there is an m-zonoset
  - intersecting exactly the $\mathcal{C}$-orthants of $\mathbb{R}^n$ at points of equal ℓ_1-norm,
  - then $\mathrm{RBM}_{n,m}$ contains a distribution with strong modes $\mathcal{C}$ and, furthermore, $\mathcal{C}$ is perfectly reconstructible.

Hyperplane arrangements

- A *hyperplane arrangement* $\mathcal{A}$ in $\mathbb{R}^n$ is a finite set of (affine) hyperplanes $\{H_i\}_{i\in[k]}$ in $\mathbb{R}^n$.
- Choose an orientation for each hyperplane,
- each vector $x \in \mathbb{R}^n$ gets a sign vector $\mathrm{sgn}_{\mathcal{A}}(x) \in \{-, 0, +\}^k$, indicating whether x lies on the negative side, inside, or on the positive side of each H_i.
- The set of all vectors in $\mathbb{R}^n$ with the same sign vector is called a *cell* of $\mathcal{A}$.

Linear threshold codes

Recall

Definition

A *linear threshold function* (LTF) with m (binary) inputs is a function

$$f\colon \{0,1\}^m \to \{-,+\}; \quad y \mapsto \operatorname{sgn}((\sum_{j\in[m]} w_j y_j) + b);$$

where $w \in \mathbb{R}^m$ is called *weight vector* and $b \in \mathbb{R}$ *bias*.

- When $f(x_1,\ldots,x_m) = \overline{f}(\overline{x_1},\ldots,\overline{x_m})$ for all $x \in \{0,1\}^m$, the LTF is *self-dual*.
- An LTF with an equal number of positive and negative points separates every input from its opposite and is self-dual.

Linear threshold codes

Definition

A subset $\mathcal{C} \subseteq \{0,1\}^n \cong \{-,+\}^n$ is an (n,m)-*linear threshold code* (LTC) if there exist n linear threshold functions $f_i \colon \{0,1\}^m \to \{0,1\}$, $i \in [n]$ with

$$\{(f_1(y), f_2(y), \ldots, f_n(y)) \in \{0,1\}^n \colon y \in \{0,1\}^m\} = \mathcal{C}.$$

- Equivalently, $\mathcal{C}$ is an (n,m)-LTC if it is the image of a down inference function of $\mathrm{RBM}_{n,m}$.
- If all f_i can be chosen self-dual, then $\mathcal{C}$ is called *homogeneous*.

LTC Example: $n = 3$ and $m = 2$

- There are only two ways to linearly separate the vertices of the 2-cube into sets of cardinality two (up to opposites): $12\overline{34}$ and $1\overline{2}3\overline{4}$.
- These are the only possible columns of a homogeneous LTC with two inputs (up to opposites).
- But the even or odd parity code is not a $(3, 2)$-LTC
- So, there does not exist a 2-zonoset with vertices in the four even, or odd, orthants of $\mathbb{R}^3$, and the $\mathrm{RBM}_{3,2}$ does not contain any distributions with four strong modes.

But there are 104 ways to linearly separate the vertices of the $n = 3$-cube. Here is a good one:

LTC Example: $n = 4$ and $m = 3$

The 8 vertices of the 3-cube are in the 8 even-parity cells of an arrangement of four hyperplanes corresponding to the $(4, 3)$-LTC with the LTFs:

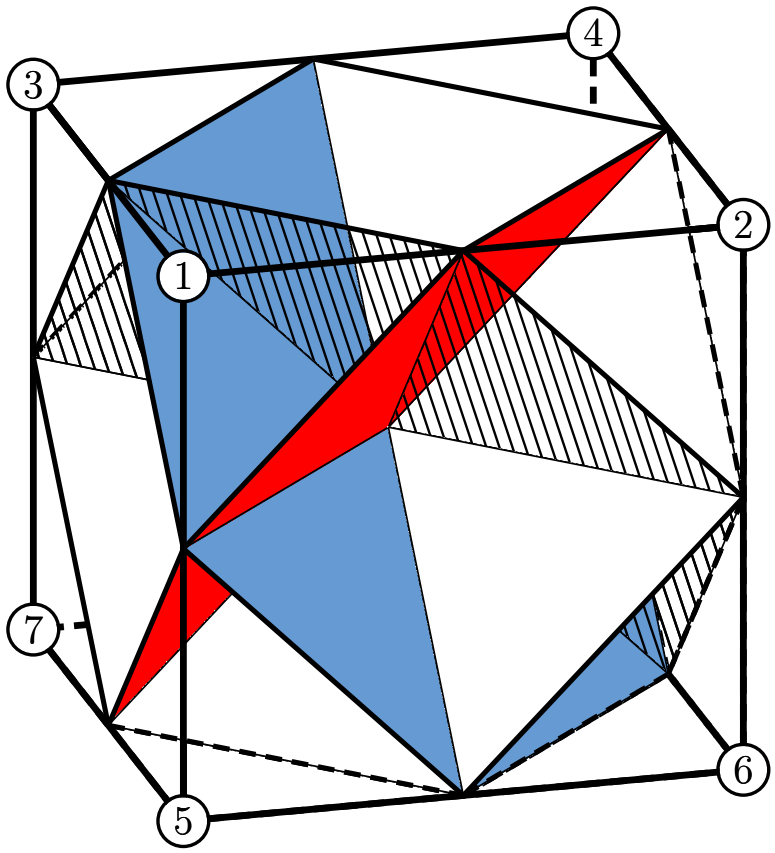
$$123\overline{4}5\overline{6}\overline{7}\overline{8},\ 12\overline{3}4\overline{5}6\overline{7}\overline{8},\ 1\overline{2}34\overline{5}\overline{6}7\overline{8},\ 1\overline{2}\overline{3}\overline{4}5\overline{6}7\overline{8}.$$

This arrangement corresponds to a 3-zonoset with points in the 8 even orthants of $\mathbb{R}^4$, realizable as:

$$w = \begin{pmatrix} -1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 \end{pmatrix}; \quad \mathcal{Z} = \frac{1}{2}\begin{pmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & -1 & -1 \\ 1 & -1 & 3 & -1 \\ -1 & 1 & 1 & -3 \\ 1 & -1 & -1 & 3 \\ -1 & 1 & -3 & 1 \\ -1 & -3 & 1 & 1 \\ -3 & -1 & -1 & -1 \end{pmatrix}.$$

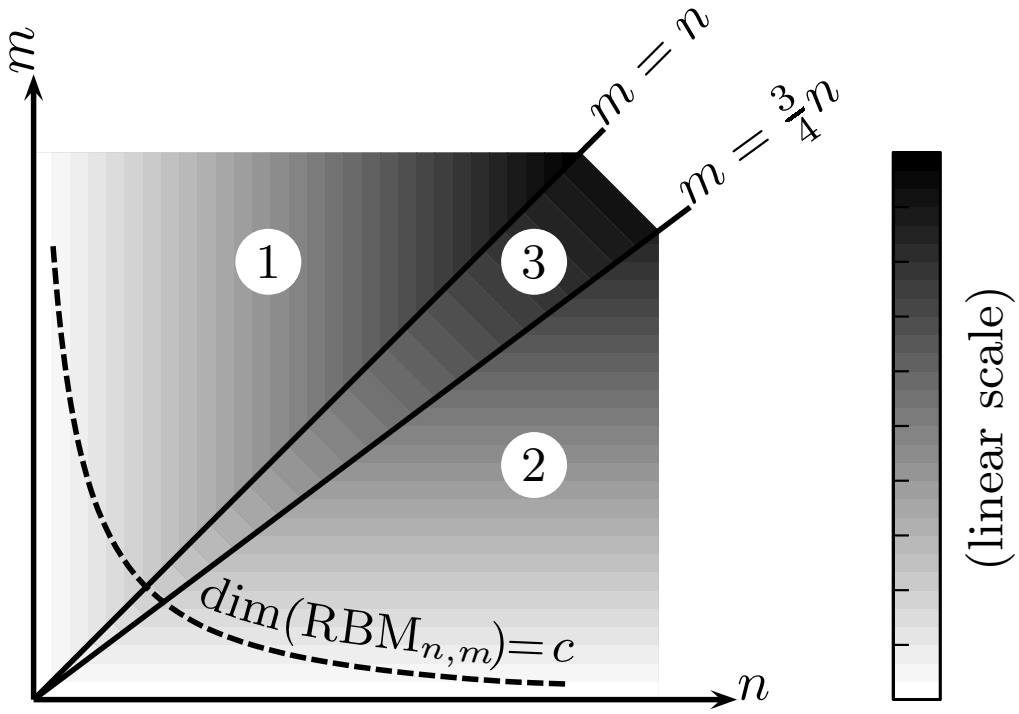
$$b = \frac{1}{2}\begin{pmatrix} 3 & 1 & 1 & 1 \end{pmatrix};$$

LTC Example: $n = 4$ and $m = 3$



These four slicings of the 3-cube can be "repeated" to obtain:

Smallest mixture of products representing an RBM



(1) $\frac{3}{4}n \leq \log_2(k) \leq n-1$

(2) $\log_2(k) = m$

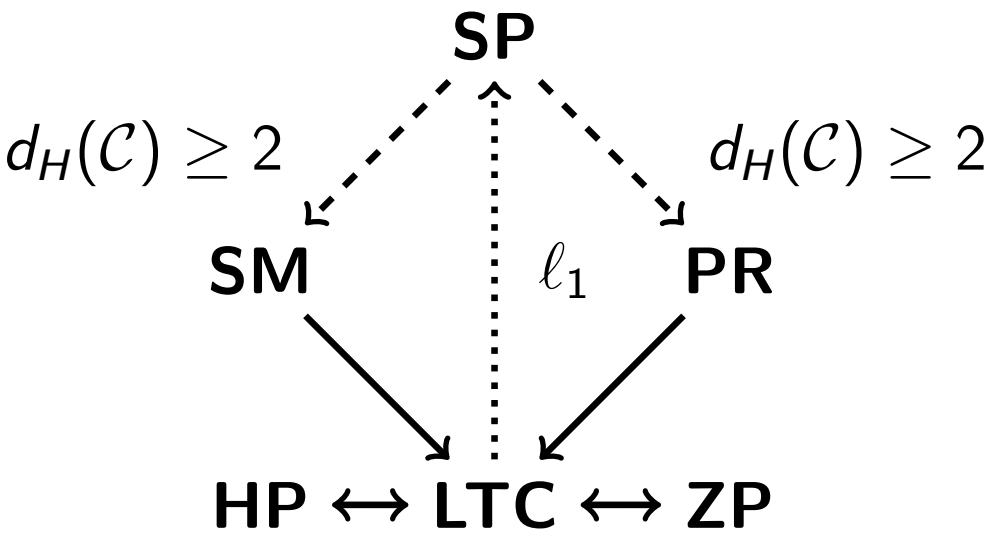
(3) $\frac{3}{4}n \leq \log_2(k) \leq m$

- Heat map of log of $k(n,m) = \min\{k' \in \mathbb{N} \colon \mathcal{M}_{n,k'} \supseteq \mathrm{RBM}_{n,m}\}$.
- To improve on our result, just find an $\mathrm{RBM}_{n,m}$ where m/n is closer to one.
- Hint: $5/6$ fails.

Six 3-parameter properties of sets of binary vectors

Let **n**, **m** be non-negative integers and $\mathcal{C}$ be a subset of $\{0,1\}^n$.

- **LTC**: The set $\mathcal{C}$ is an (n, m)-**l**inear **t**hreshold **c**ode, i.e., the image of n linear threshold functions with m inputs.
- **HP**: There exists an arrangement $\mathcal{A}$ of n **h**yper**p**lanes in $\mathbb{R}^m$ such that the vertices of the m-dimensional unit cube intersect exactly the $\mathcal{C}$-cells of $\mathcal{A}$.
- **ZP**: There is an affine image of the vertices of an m-cube (m-**z***onoset*) in $\mathbb{R}^n$ which intersects exactly the $\mathcal{C}$-orthants of $\mathbb{R}^n$.
- **SM**: An RBM with n visible and m hidden nodes can represent a distribution with set of **s**trong **m**odes $\mathcal{C}$.
- **PR**: The set $\mathcal{C}$ is the set of **p**erfectly **r**econstructible inputs of an RBM with n visible and m hidden nodes.
- **SP**: An RBM with n visible and m hidden nodes can represent a distribution which is **s**trictly **p**ositive on $\mathcal{C}$ and zero elsewhere.



Theorem

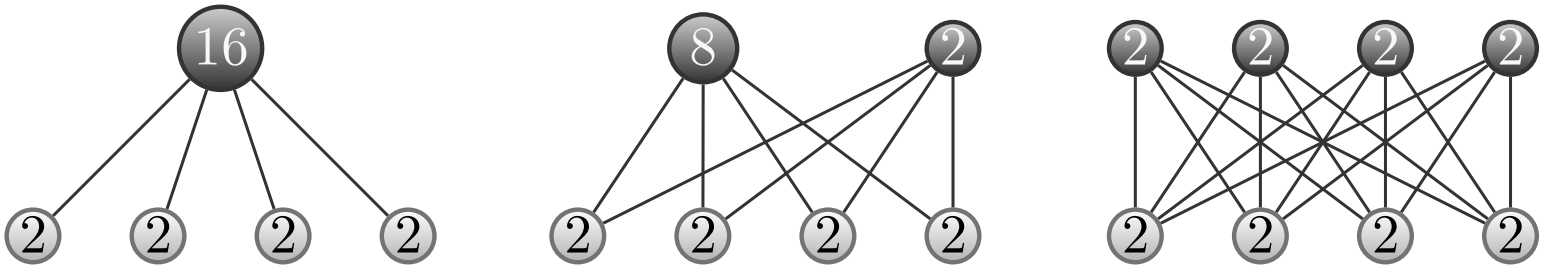
Fix integers n and m. For any $\mathcal{C} \subset \{0,1\}^n$, the following hold.

1. *The properties* **LTC**, **HP,** *and* **ZP** *are equivalent.*
2. *If $\mathcal{C}$ satisfies* **PR** *or* **SM**, *then it is contained in an* **LTC** *set.*
3. *If the vectors in $\mathcal{C}$ are at least Hamming distance 2 apart, then* **SP** *implies both* **SM** *and* **PR**.
4. *If the vectors in $\mathcal{C}$ are at least Hamming distance 2 apart and $\mathcal{C}$ satisfies an ℓ_1 property, then* **LTC** *implies* **SP**.

Discrete Restricted Boltzmann Machines

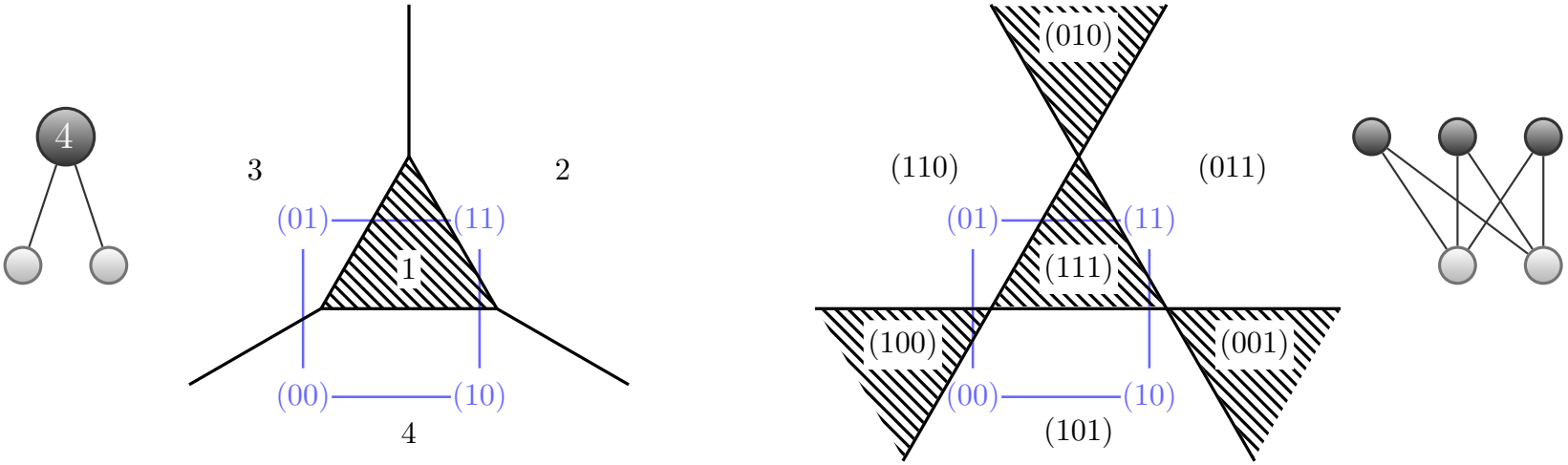
- Visible variables in RBMs are restricted to be binary; sometimes want to model multi-valued variables.
- In practice, Boltzmann machines often have sparsity restrictions: some hidden variables are treated as mutually exclusive or nearly so by penalizing multiple activation.
- We can deal with both directly by allowing both hidden and visible variables to take on more than two variables.
- e.g. in a topic model, some topics are mutually exclusive (belong to same hidden node) while others can be combined freely.

- A discrete RBM interpolates between naïve Bayes models and binary RBMs:



naïve Bayes binary RBM

- Linear vs. exponential number of inference regions:



Naïve Bayes model

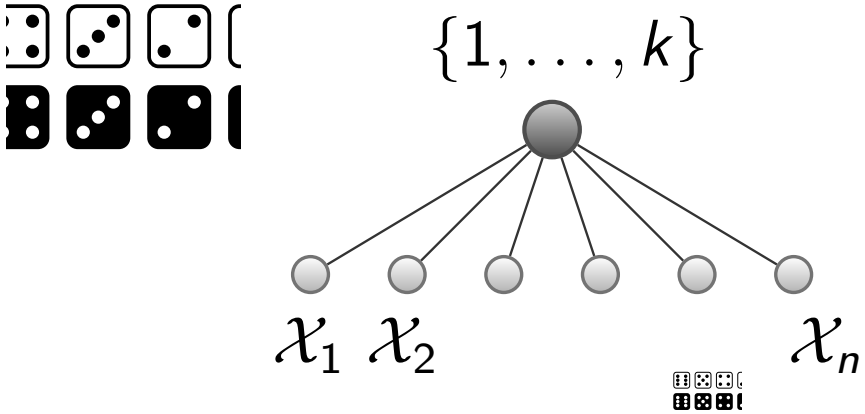
The set of all mixtures of k product distributions

$$p(x_1, \ldots, x_n) = \prod_{i \in [n]} p_i(x_i) = \frac{1}{Z(\vartheta)} \exp(\vartheta^\top \mathbf{x})$$

with parameters

$$\vartheta = \Theta A_y^{(\mathcal{Y})} \quad \text{for all } y \in \mathcal{Y} = \{1, 2, \ldots, k\},$$

whereby the columns $A_y^{(\mathcal{Y})}$ of the sufficient statistics matrix $A^{(\mathcal{Y})}$ are the vertices of a $(k-1)$-simplex.



Binary restricted Boltzmann machine

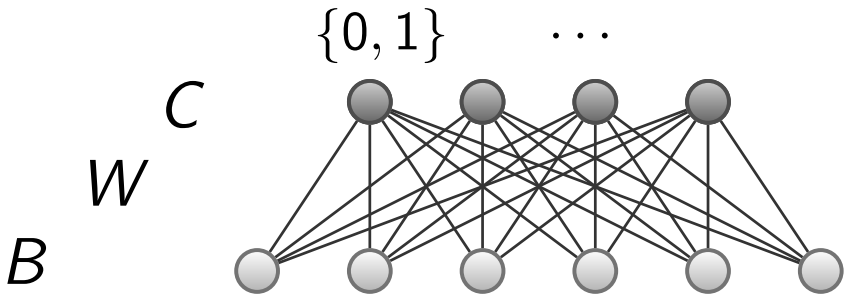
A set of *distributed* mixtures of exponentially many product distributions

$$p(x_1, \ldots, x_n) = \prod_{i \in [n]} p_i(x_i) = \frac{1}{Z(\vartheta)} \exp(\vartheta^\top \boldsymbol{x})$$

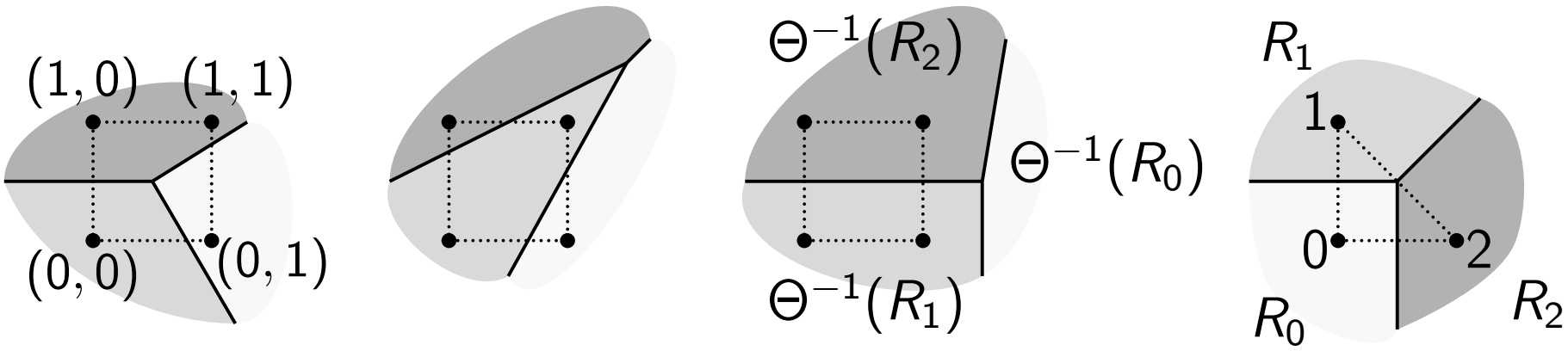
with parameters

$$\vartheta = \Theta A_y^{(\mathcal{Y})} \quad \text{for all } y \in \{0,1\}^m,$$

whereby the columns $A_y^{(\mathcal{Y})}$ of the sufficient statistics matrix $A^{(\mathcal{Y})}$ are the vertices of an m-cube.



- Binary RBMs can be described in terms of projections of cubes into hyperplane arrangements
  - The cubes are the convex supports of binary independence models
  - The hyperplanes separate the preferred inputs of binary units.
- Discrete RBMs in turn can be described in terms of projections of products of simplices and normal fans of products of simplices



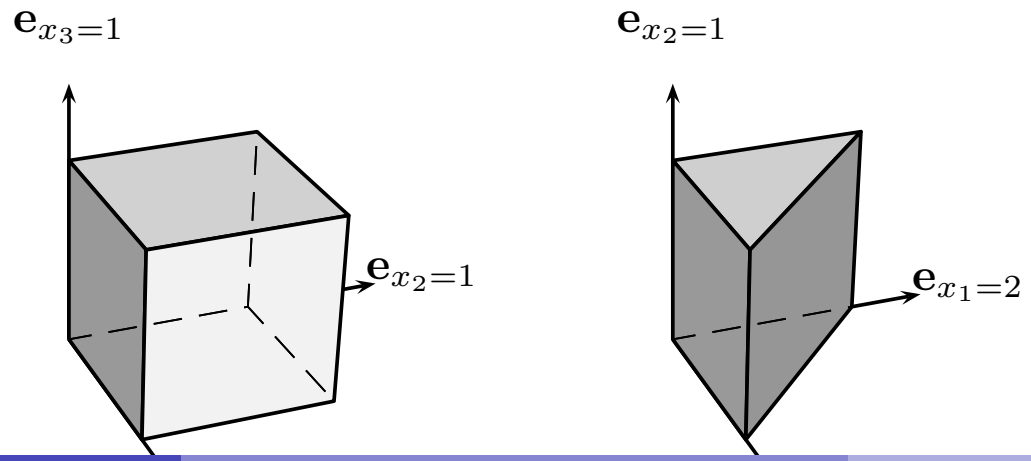
Product of Experts

A discrete RBM model is a renormalized entry-wise product of naïve Bayes models; one for each hidden unit:

$$\mathrm{RBM}_{\mathcal{X},\mathcal{Y}} = \mathrm{NB}_{\mathcal{X},\mathcal{Y}_1} \circ \cdots \circ \mathrm{NB}_{\mathcal{X},\mathcal{Y}_m} \ .$$

Distributed Mixtures

The set of conditionals $p_\Theta(x|y)$, $y \in \mathcal{Y}$ is the set of product distributions with parameters $\Theta^\top A_y^{(\mathcal{Y})}$, $y \in \mathcal{Y}$, equal to a linear projection of the vertices of the Cartesian product of simplices $\Delta(\mathcal{Y}_1) \times \cdots \times \Delta(\mathcal{Y}_m)$.



Representational Power

We describe explicit classes of distributions that can be represented by an RBM depending on the state spaces of its variables:

Theorem

The model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *contains any mixture distribution*

$$p = \sum_{j \in \{0,1,\ldots,m\}} \lambda_j \, p^{(j)},$$

where $p^{(j)}$ *is any mixture of* $|\mathcal{Y}_j| - 1$ *product distributions and all* $p^{(j)}$ *have disjoint supports for all* $j \in [m]$.

- This describes how the rigidity of the mixtures decreases when the hidden variables have larger state spaces, and RBM becomes more like NB.

Approximation Errors

Theorem

The Kullback-Leibler divergence from any distribution p to a ML projection in the model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *is bounded by*

$$D(p\|p_{\mathrm{RBM}}) \;\leq\; \log|\mathcal{X}| - \log\sum_{j\in[m]}(|\mathcal{Y}_j|-1)$$

- The maximal approximation error increases at most logarithmically with the number of visible states, and
- decreases at least logarithmically in sum of hidden state numbers

Corollary

The model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *is a universal approximator of dist. on* $\mathcal{X}$ *if*

$$1+\sum_{j\in[m]}(|\mathcal{Y}_j|-1) \;\geq\; |\mathcal{X}|/\max_{i\in[n]}|\mathcal{X}_i|$$

Dimension

The "expected" dimension of a model on $\mathcal{X}$ with hidden variables is equal to the number of parameters, or to $|\mathcal{X}| - 1$.

- In some cases discrete RBMs do not have the expected dimension; when the naïve Bayes factors don't.
- In many cases they do:

Theorem

The model $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ *has the dimension*

- *equal to the number of parameters, whenever* $\mathcal{X}$ *contains* m *disjoint Hamming balls of radii* $2(|\mathcal{Y}_j| - 1) - 1$, $j \in [m]$
- *equal to* $|\mathcal{X}| - 1$, *when* m *radius-one Hamming balls cover* $\mathcal{X}$.

Tropical Model

The tropical model $\mathrm{RBM}^{\mathrm{tropical}}_{\mathcal{X},\mathcal{Y}}$ is the image of the map

$$\begin{aligned}\Phi(x;\theta) &= \max\{\langle \theta, A^{(\mathcal{X},\mathcal{Y})}_{(x,y)}\rangle : y \in \mathcal{Y}\} \quad \text{for all } x \in \mathcal{X}, \theta \in \mathbb{R}^d, \\ &= \langle \theta, A^{(\mathcal{X})}_{x} \otimes A^{(\mathcal{Y})}_{y_\theta(x)}\rangle\end{aligned}$$

The dimension of $\mathrm{RBM}_{\mathcal{X},\mathcal{Y}}$ is lower bounded by the rank of the matrix

$$\left(A^{(\mathcal{X})}_{x} \otimes A^{(\mathcal{Y})}_{y_\theta(x)}\right)_{x,y_\theta}$$

and can be estimated by solving linear optimization problems, which lead to purely combinatorial problems (such as error correcting codes).

Conclusions

- RBMs and discrete RBMS are marginalizations of models where the matrix of sufficent statistics is the Kronecker (tensor) product of the sufficent statistics matrices for the hidden and visible variables (in the RBM case, Segre matrices).
- Generalizing results to such arbitrary Kronecker products yields solution to the dimension problem for such models.
- In particular, the binary RBM always has the expected dimension.

Thank you