

Geometry of the Restricted Boltzmann Machine

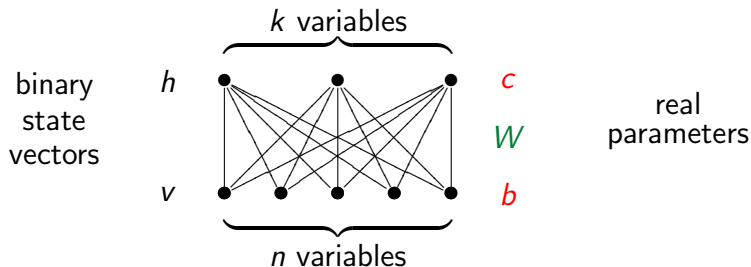
Jason Morton

Penn State

AMS Sectional Meeting
March 27, 2010

Joint work with M.A. Cueto and B. Sturmfels of U.C. Berkeley.

Graphical model on a bipartite graph



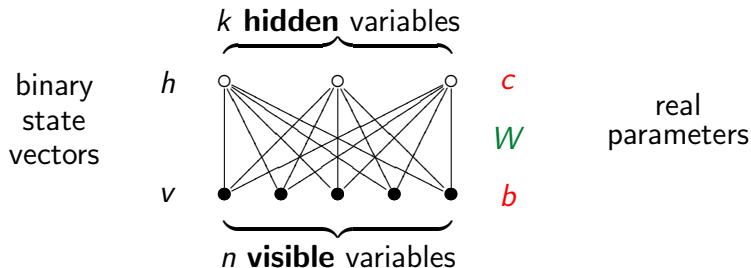
Unnormalized **potential** is built from **node** and **edge** parameters

$$\psi(v, h) = \exp(h^\top W v + b^\top v + c^\top h).$$

The probability distribution on the binary random variables is

$$p(v, h) = \frac{1}{Z} \cdot \psi(v, h), \quad Z = \sum_{v, h} \psi(v, h).$$

Restricted Boltzmann machines



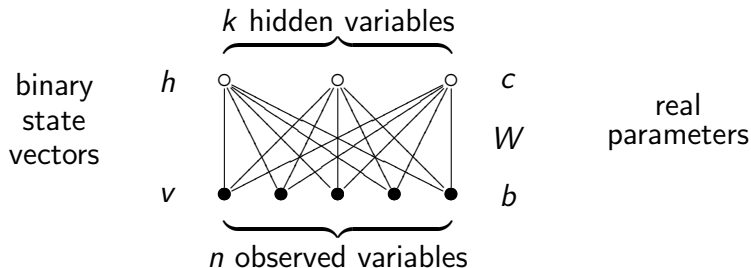
Unnormalized fully-observed **potential** is

$$\psi(v, h) = \exp(h^\top W v + b^\top v + c^\top h).$$

The probability distribution on the visible random variables is

$$p(v) = \frac{1}{Z} \cdot \sum_{h \in \{0,1\}^k} \psi(v, h), \quad Z = \sum_{v, h} \psi(v, h).$$

Restricted Boltzmann machines



- The *restricted Boltzmann machine* (RBM) is the undirected graphical model for binary random variables thus specified.
- Denote by M_n^k the set of joint distributions as $b \in \mathbb{R}^n$, $c \in \mathbb{R}^k$, $W \in \mathbb{R}^{k \times n}$ vary.
- M_n^k is a subset of the probability simplex $\Delta_{2^n - 1}$.

Modeling with RBMs

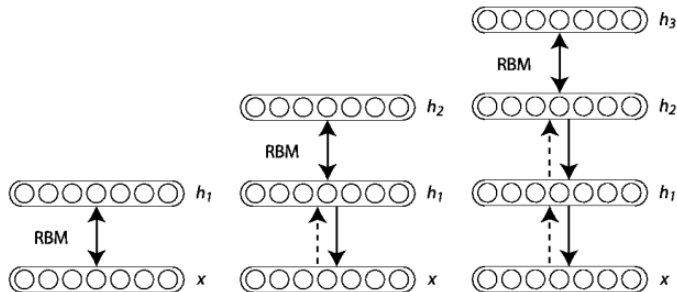
- Used to model distributions on the visible nodes, $\{0, 1\}^n$
- Idea: observe many independent draws from unknown distribution, fit parameters with maximum likelihood
- Training is quite hard
- Hinton-Osindero-Teh 2006: Contrastive divergence (SGA on v^*)

$$\frac{\partial \log p_W(v^*)}{\partial w_{ij}} = \mathbb{E}_{p_{W, v^*}}[v_i^* h_j] - \mathbb{E}_{p_W}[v^i h_j]$$

- $\mathbb{E}_{p_{W, v^*}}[v_i^* h_j] = 0$ if $v_i^* = 0$, or $1/(1 + \exp(-\sum_i v_i^* w_{ij}))$
- $\mathbb{E}_{p_W}[v^i h_j]$ involves a partition function, estimated by MCMC
- Is the model even identifiable?

Why restricted Boltzmann machines?

- The RBM has become extremely popular in machine learning due to its role as the building block of the deep belief network.
- Deep belief networks aim to learn feature hierarchies to automatically find high-level representations of high-dimensional data (see [Hinton 2006]).



Many new applications in general machine learning problems including object recognition and dimensionality reduction.

Stacked RBMs: flexible generative models



Stacked RBMs: flexible generative models



[Lee, Grosse, Ranganath, Ng 2009]

Representational power of RBMs

- Promising experiments—but the scope and basic properties of DBNs poorly understood.
- Le Roux and Bengio 2008: any distribution with support on r visible states may be arbitrarily well approximated provided there are at least $r + 1$ hidden nodes.
 - ▶ So all distributions can be approximated with $k = 2^n + 1$ hidden nodes
 - ▶ vs. $k = \frac{2^n - (n+1)}{n+1}$ best case
- How well can we do? **Is the restricted Boltzmann machine identifiable?**
- The dimension of the fully observed binary graphical model on $K_{k,n}$ is equal to $nk + n + k$, the number of nodes plus the number of edges.
- Marginalizing/hiding nodes is a **big linear projection**; does it preserve dimension?

Representational power of RBMs

Conjecture

The restricted Boltzmann machine has the expected dimension, i.e. M_n^k is a semialgebraic set of dimension $\min\{nk + n + k, 2^n - 1\}$ in Δ_{2^n-1} .

We can show many special cases and the following general result:

Theorem

The restricted Boltzmann machine has the expected dimension

- $nk + n + k$ when $k < 2^{n - \lceil \log_2(n+1) \rceil}$
 - $\min\{nk + n + k, 2^n - 1\}$ when $k = 2^{n - \lceil \log_2(n+1) \rceil}$ and
 - $2^n - 1$ when $k \geq 2^{n - \lfloor \log_2(n+1) \rfloor}$.
- Covers most cases of restricted Boltzmann machines in practice, as those generally satisfy $k \leq 2^{n - \lceil \log_2(n+1) \rceil}$.

Four geometric objects on the way to this result.

Player 1: RBM Model M_n^k

The **RBM model** M_n^k is the set of all probability distributions $(p_v)_{v \in \{0,1\}^n}$ that can be written as

$$p_v = \frac{1}{Z} \cdot \sum_{h \in \{0,1\}^k} \exp(h^\top Wv + b^\top v + c^\top h).$$

for some choice of $W \in \mathbb{R}^{k \times n}$, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}^k$. There are $nk + n + k$ model parameters; is $\dim(M_n^k) = \min\{nk + n + k, 2^n - 1\}$?

Proposition

The RBM model M_n^k is a semialgebraic subset of the simplex Δ_{2^n-1} .

Algebraic statistics and graphical models

- Now consider the structure of the model as an algebraic variety
 - ▶ For this and other discrete models, the space of probability distributions is the image of a polynomial parameterization.
 - ▶ Includes graphical models for both Gaussian and discrete random variables.
- Studying the **Zariski closure** and **tropical variety** gives us a lot of information about the space of probability distributions modeled

Player 2: RBM Variety V_n^k

The RBM model M_n^k is the image of the map $\mathbb{R}_{>0}^{nk+k+n} \rightarrow \Delta_{2^n-1}$ whose 2^n coordinates are

$$p_v = \frac{1}{Z} \beta_1^{v_1} \beta_2^{v_2} \cdots \beta_n^{v_n} \prod_{i=1}^k (1 + \gamma_i \omega_{i1}^{v_1} \omega_{i2}^{v_2} \cdots \omega_{in}^{v_n}).$$

When faced with a complicated semialgebraic set arising in applications, it is useful to simplify by disregarding inequalities, and to replace \mathbb{R} with \mathbb{C} .

The *RBM variety* V_n^k is the irreducible variety in the complex projective space \mathbb{P}^{2^n-1} obtained by taking the Zariski closure of the RBM model $M_n^k \subset \Delta_{2^n-1}$.

Hadamard Products of Varieties

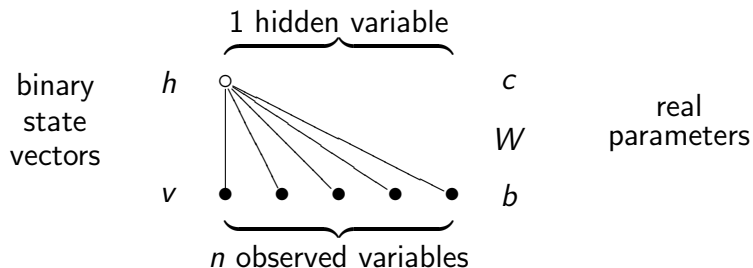
Given two projective varieties X and Y in \mathbb{P}^m , their *Hadamard product* $X * Y$ is the closure of the image of

$$X \times Y \dashrightarrow \mathbb{P}^m, (x, y) \mapsto (x_0y_0 : x_1y_1 : \dots : x_my_m).$$

We also define *Hadamard powers* $X^{[k]} = X * X^{[k-1]}$. Here V_n^1 , $V_n^1 * V_n^1, \dots$

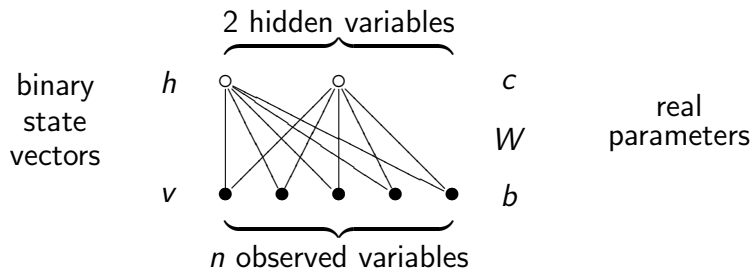
Hadamard Products of Varieties

One hidden variable: V_n^1



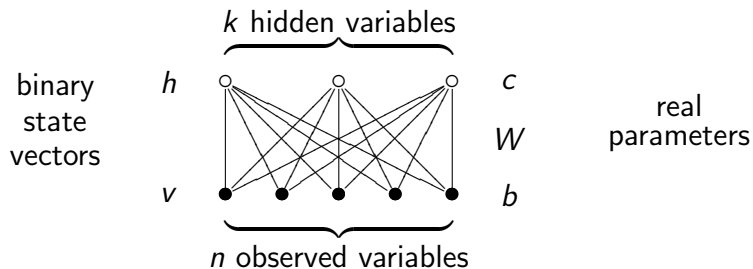
Hadamard Products of Varieties

Two hidden variables: $V_n^1 * V_n^1$



Hadamard Products of Varieties

k hidden variables: $(V_n^1)^{[k]}$



Hadamard Products of Varieties

Given two projective varieties X and Y in \mathbb{P}^m , their *Hadamard product* $X * Y$ is the closure of the image of

$$X \times Y \dashrightarrow \mathbb{P}^m, (x, y) \mapsto (x_0 y_0 : x_1 y_1 : \dots : x_m y_m).$$

We also define *Hadamard powers* $X^{[k]} = X * X^{[k-1]}$.

If M is a subset of the simplex Δ_{m-1} then $M^{[k]}$ is also defined by componentwise multiplication followed by rescaling so that the coordinates sum to one. This is compatible with taking Zariski closure: $\overline{M^{[k]}} = \overline{M}^{[k]}$

Lemma

RBM variety and RBM model factor as

$$V_n^k = (V_n^1)^{[k]} \quad \text{and} \quad M_n^k = (M_n^1)^{[k]}.$$

Secant lines of the Segre Variety

The first RBM variety V_n^1 coincides with the **first secant variety of the Segre embedding of $(\mathbb{P}^1)^n$** into \mathbb{P}^{2^n-1} . The first tropical RBM variety TV_n^1 is the tropicalization of that secant variety.

Theorem (Landsberg-Manivel)

The projective variety $V_n^1 \subset \mathbb{P}^{2^n-1}$ is cut out by the 3×3 -minors of all flattenings of the n -dimensional tensor $(p_\nu)_{\nu \in \{0,1\}^n}$.

Zwiernik and Smith (2009) describe additional inequalities that characterize the model $M_n^1 \subsetneq V_n^1$ such as $\sigma_{12}\sigma_{13}\sigma_{23} \geq 0$.

Player 3: Tropical RBM Model

- *Tropical mathematics* is predicated on the idea that $\log(\exp(x) + \exp(y))$ is approximately equal to $\max(x, y)$ when x and y are quantities of different scale.
- The process of passing from ordinary arithmetic to the max-plus algebra is known as *tropicalization*.
- The tropicalization Φ of our given morphism p is the map $\Phi : \mathbb{R}^{nk+n+k} \rightarrow \mathbb{TP}^{2^n-1} = \mathbb{R}^{2^n}/\mathbb{R}(1, 1, \dots, 1)$ whose 2^n coordinates are the tropical polynomials

$$q_v = \max\{h^\top Wv + b^\top v + c^\top h : h \in \{0, 1\}^k\}$$

- This yields a piecewise-linear concave function $\mathbb{R}^{nk+n+k} \rightarrow \mathbb{R}$ on the space of model parameters (W, b, c) .

Its **image** TM_n^k is called the *tropical RBM model*.

Parametric Inference

The tropical model TM_n^k is a geometric object built from the graph that organizes the space of **inference functions** which the model can compute [Pachter and Sturmfels 2004].

Given an RBM with known parameters, we infer which value \hat{h} of the hidden data maximizes $\text{Prob}(h \mid v)$, using it to classify or feed to another RBM. Each parameter choice $\theta = (b, W, c)$ defines an *inference function*

$$l_\theta : \{0, 1\}^n \rightarrow \{0, 1\}^k, v \mapsto \hat{h}.$$

The value $l_\theta(v)$ is the hidden string $\hat{h} \in \{0, 1\}^k$ that attains the maximum in the tropical polynomial

$$q_v = b^\top v + \max_{h \in \{0, 1\}^k} \{h^\top Wv + c^\top h\}.$$

Linear Threshold Functions

The inference functions for the RBM model M_n^k are precisely those Boolean functions $\{0, 1\}^n \rightarrow \{0, 1\}^k$ whose k coordinates are **linear threshold functions**, i.e. slicings of the cube. The number $\lambda(n)$ of linear threshold functions $\{0, 1\}^n \rightarrow \{0, 1\}$ satisfies [Ojha 2000]

$$2^{\binom{n}{2}+16} < \lambda(n) \leq 2^{n^2}.$$

Corollary

The RBM model M_n^k has $\lambda(n)^k = 2^{\Theta(kn^2)}$ inference functions.

[Sloane-Plouffe] lists $\lambda(n)$ for $n = 1, 2, \dots, 8$:

4, 14, 104, 1882, 94572, 15028134, 8378070864, ...

But only $64(nk + n + k)$ bits of parameter

Player 4: Tropical RBM Variety

- Finally, we define the **tropical RBM variety** TV_n^k to be the tropicalization of the RBM variety V_n^k .
- The **tropical hypersurface** $\mathcal{T}(f)$ is the union of all codimension one cones in the normal fan of the Newton polytope of f .
- The tropical variety TV_n^k is the intersection in \mathbb{TP}^{2^n-1} of all the tropical hypersurfaces $\mathcal{T}(f)$ where f runs over *all* polynomials that vanish on V_n^k (or on M_n^k).
- If the homogeneous prime ideal of the variety V_n^k were known then the tropical variety TV_n^k could in theory be computed using e.g. Gfan.
- Even for small instances, ideal computation is very hard. [Cueto and Yu 2008]: V_4^2 is a hypersurface of degree 110 in \mathbb{P}^{15} , and the defining irreducible polynomial is unknown.

Player 4: Tropical RBM Variety

- Recall: The projective variety $V_n^1 \subset \mathbb{P}^{2^n-1}$ is cut out by the 3×3 -minors of all flattenings of the n -dimensional tensor $(p_v)_{v \in \{0,1\}^n}$.
- Do the 3×3 flattening minors from a tropical basis? Do they cut out TV_k^1 tropically?
- No $n = 4$

The Four Players

$$\begin{aligned} \text{Classical : } & M_n^k \subset V_n^k \subset \mathbb{P}^{2^n-1} \\ \text{Tropical : } & TM_n^k \subset TV_n^k \subset \mathbb{TP}^{2^n-1} \end{aligned}$$

The dimensions of these four objects satisfy

$$\begin{aligned} \dim(TM_n^k) \leq \dim(TV_n^k) = \dim(V_n^k) = \\ \dim(M_n^k) \leq \min\{nk + n + k, 2^n - 1\}. \end{aligned}$$

Conjecture

The tropical RBM model has the expected dimension, i.e. TM_n^k is a polyhedral fan of dimension $\min\{nk + n + k, 2^n - 1\}$ in \mathbb{TP}^{2^n-1} .

Theorem

Conjectures are true when $k \leq 2^{n - \lceil \log_2(n+1) \rceil}$ and when $k \geq 2^{n - \lfloor \log_2(n+1) \rfloor}$.

Regions of linearity in the tropical morphism

Recall the tropical morphism $\Phi : \Theta = \mathbb{R}^{nk+n+k} \rightarrow \mathbb{TP}^{2^n-1}$ with

$$q_v = b^\top v + \max_{h \in \{0,1\}^k} \{h^\top Wv + c^\top h\}.$$

- Φ is a **piecewise linear** map with image TM_n^k .
- At a point $\theta = (c, W, b)$, the map is defined by a matrix $\mathcal{A}_\theta \in \mathbb{R}^{2^n \times (nk+n+k)}$
- Then $\dim(TM_n^k) \geq \text{rank}(\mathcal{A}_\theta)$.
- So, we just need to **determine** \mathcal{A}_θ and **lower bound its rank**.

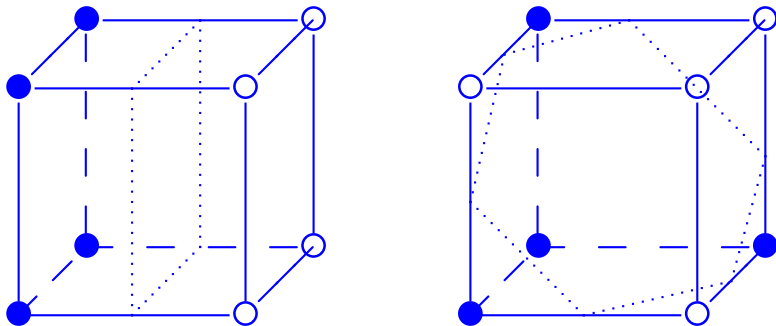
Slicings of the n -cube

- A subset C of the vertices of the n -cube is a *slicing* if there exists a hyperplane that has the vertices in C on the positive side and the remaining vertices of the n -cube on the other side.
- Idea: a slicing
 - ▶ corresponds to a single hidden node h_i and the associated parameters;
 - ▶ represents a division of the possible visible states $\{0, 1\}^n$ into those where the inferred state for h_i is 0 and those where it is 1, and
 - ▶ the division is linear since model is loglinear.

Begin with $k = 1, n = 3 \dots$

$$q_v = b^\top v + \max\{0, \omega v + c\}$$

Slicings of the n -cube and $\dim(TM_n^1)$



Partitions of $\{0, 1\}^3$ defining non-empty cones where the tropical morphism is linear. Slicing on the right represents a cone in the parameter space whose image is full-dimensional, while the one on the left does not.

Proposition

The first tropical RBM model TM_n^1 has the expected dimension $2n + 1$.

Dimension Formula for TM_n^k

Let A denote the $2^n \times n$ matrix with rows $\{0, 1\}^n$. For any slicing C of the n -cube, let A_C be the $2^n \times (n+1)$ matrix whose rows indexed by the vertices v in C are $(1, v) \in \{0, 1\}^{n+1}$ and whose other rows are all zero.

Theorem

The dimension of the tropical RBM model TM_n^k equals the maximum rank of any $2^n \times (nk + n + k)$ matrix of the form

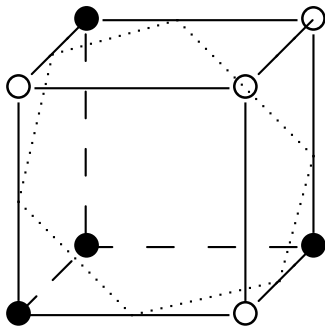
$$\mathcal{A} = (A \mid A_{C_1} \mid A_{C_2} \mid \cdots \mid A_{C_k}),$$

where C_1, C_2, \dots, C_k are k slicings of the n -cube.

Now we apply some [coding theory](#) to bound this maximum rank.

Binary codes and covering codes

Roughly: codes are families of slicings. If you can pack in k disjoint Hamming neighborhoods ($nk + n + k$), or cover everything with k Hamming neighborhoods ($2^n - 1$), the expected dimension is obtained.



Binary codes and covering codes

Roughly: codes are families of slicings. If you can pack in k disjoint Hamming neighborhoods ($nk + n + k$), or cover everything with k Hamming neighborhoods ($2^n - 1$), the expected dimension is obtained. Thus we are interested in:

- Lower bound on $A_2(n, 3)$, the size (number of codewords) of the **largest** binary code on n bits with each pair of codewords at least Hamming distance (number of bit flips) 3 apart.
- Upper bound on $K_2(n, 1)$, the size of the **smallest** n -bit covering code. $K_2(n, 1)$ is the least number of codewords such that every string of n bits is within Hamming distance 1 of some codeword.

Corollary

- $\dim TM_n^k = nk + n + k$ for $k < A_2(n, 3)$,
- $\dim TM_n^k = \min\{nk + n + k, 2^n - 1\}$ for $k = A_2(n, 3)$,
- $\dim TM_n^k = 2^n - 1$ for $k \geq K_2(n, 1)$.

Coding theory bounds

The computation of $A_2(n, 3)$ and $K_2(n, 1)$, both in general and for specific values of n , has been an active area of research since the 1950s. In addition to many known results for specific values of n , the following bounds can be obtained.

Proposition

For binary codes with $n \geq 3$, the Varshamov bound holds:

$$A_2(n, 3) \geq 2^{n - \lceil \log_2(n+1) \rceil}.$$

For covering codes, the following inequality holds:

$$K_2(n, 1) \leq 2^{n - \lfloor \log_2(n+1) \rfloor}.$$

For $n = 2^\ell - 1$ with $\ell \geq 3$ (Hamming codes), we have equality $A_2(n, 3) = K_2(n, 1) = 2^{2^\ell - \ell - 1}$.

Coding theory bounds

Corollary

The coding theory argument leads to the following bounds:

- *If $k < 2^{n - \lceil \log_2(n+1) \rceil}$, then $\dim TM_n^k = nk + n + k$.*
- *If $k = 2^{n - \lceil \log_2(n+1) \rceil}$, then $\dim TM_n^k = \min\{nk + n + k, 2^n - 1\}$.*
- *If $k \geq 2^{n - \lceil \log_2(n+1) \rceil}$, then $\dim TM_n^k = 2^n - 1$.*

End
morton@math.psu.edu