Geometry of Boltzmann Machines

Guido Montúfar
Max Planck Institute for Mathematics in the Sciences,
Leipzig

Talk at IGAIA IV, June 17, 2016

On the occasion of Shun-ichi Amari's 80th birthday

MAX-PLANCK-GESELLSCHAFT

Max Planck Institute for Mathematics in the Sciences

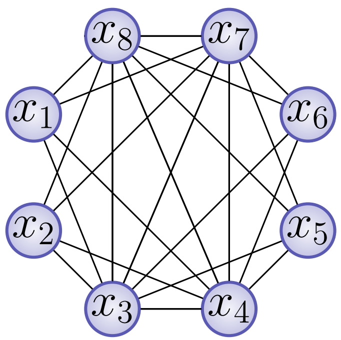
- Boltzmann Machines
- Geometric Perspectives
- Universal Approximation (new results)
- Dimension (new results)

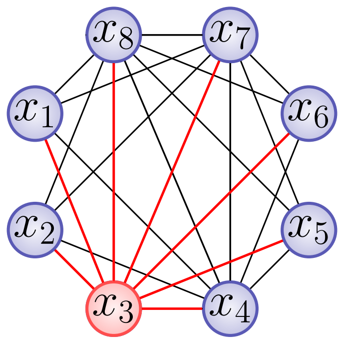
Boltzmann Machines

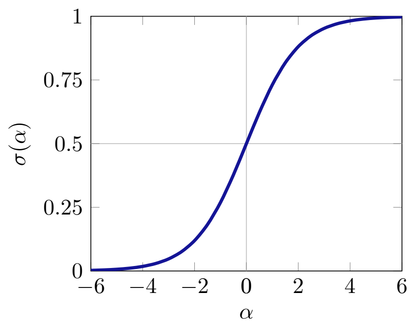
A Boltzmann machine is a network of stochastic units.
It defines a set of probability vectors

$$p_\theta(x) = \exp\left(\sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j - \psi(\theta)\right), \qquad x \in \{0,1\}^N,$$

for all $\theta \in \mathbb{R}^d$.







[Ackley, Hinton, Sejnowski '85] [Geman & Geman '84]

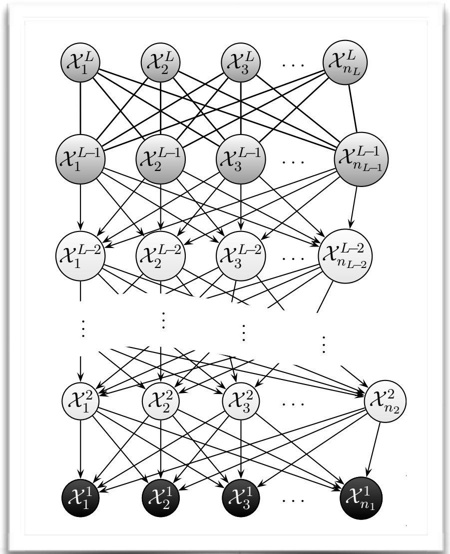
Boltzmann Machines

Generative Models

Learning Representations

Learning Modules
for Deep Belief Networks

Classification

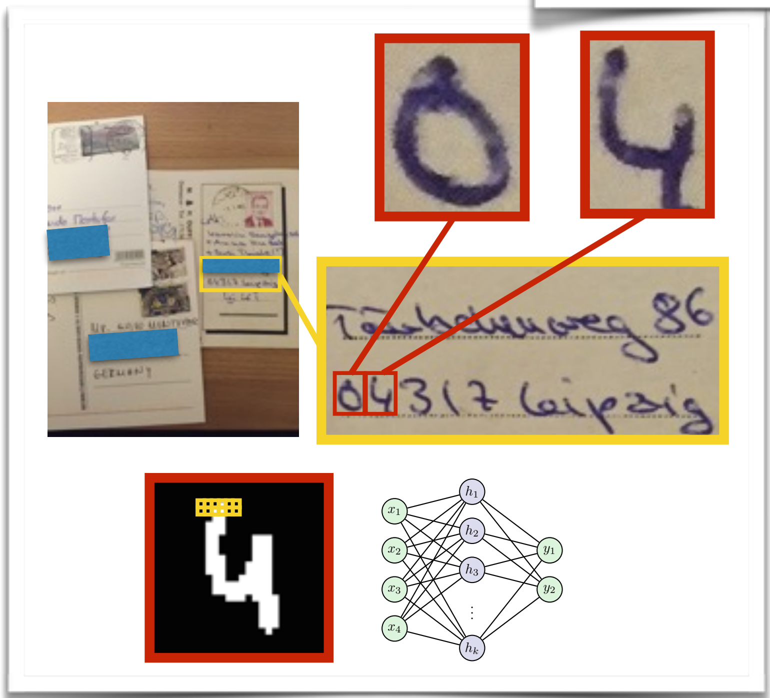


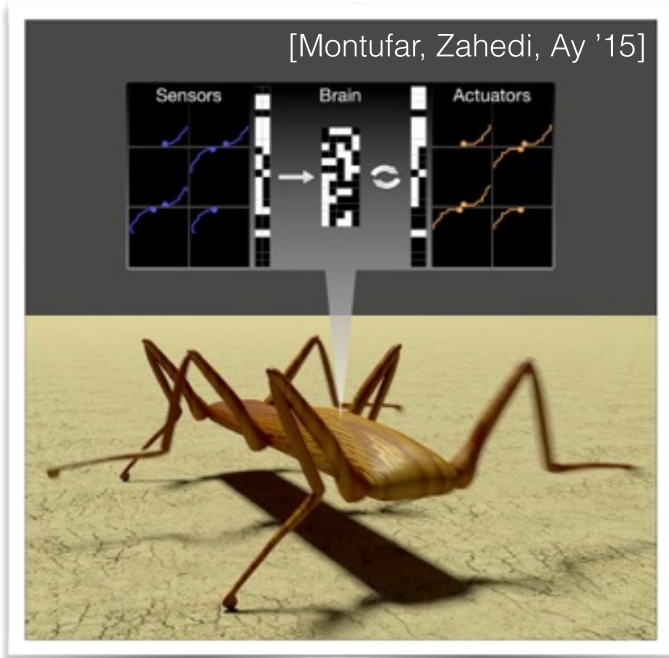
Modeling Temporal Sequences

Structured Output Prediction

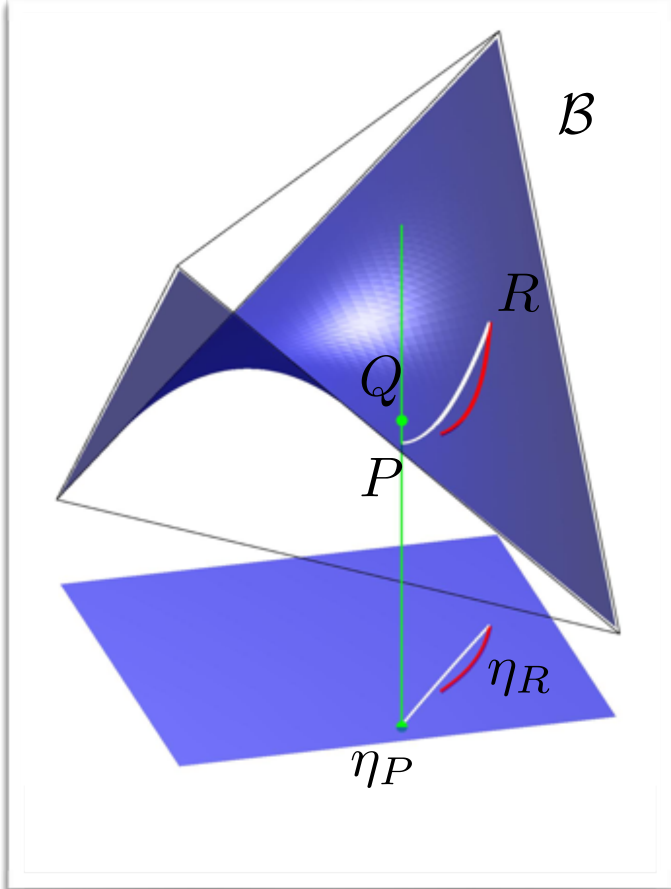
Recommender Systems

Stochastic Controller





Information Geometric Perspectives



$$\eta = \nabla\psi(\theta)$$

$$\Delta\theta = \epsilon G^{-1}(\eta_Q - \eta_R)$$

Without hidden units

$$p_\theta(x) = \exp\left(\sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j - \psi(\theta)\right)$$

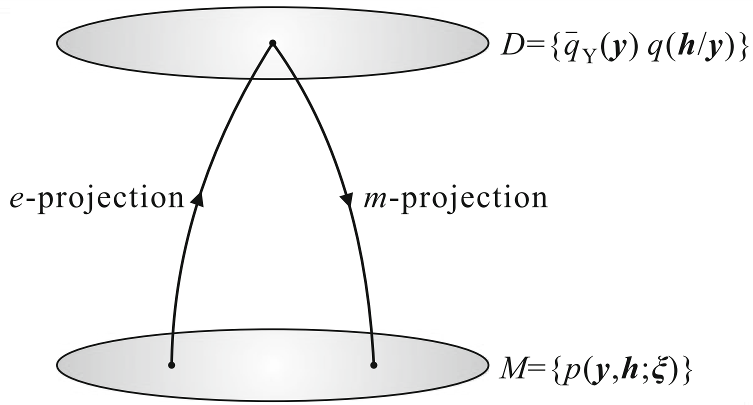
- The Boltzmann machine defines an e-linear manifold
- MLE is the unique m-projection of the target distribution to this manifold
- Natural gradient learning trajectory is the m-geodesic to the MLE
- Stochastic interpretation of natural parameters

[Amari, Kurata, Nagaoka '92]

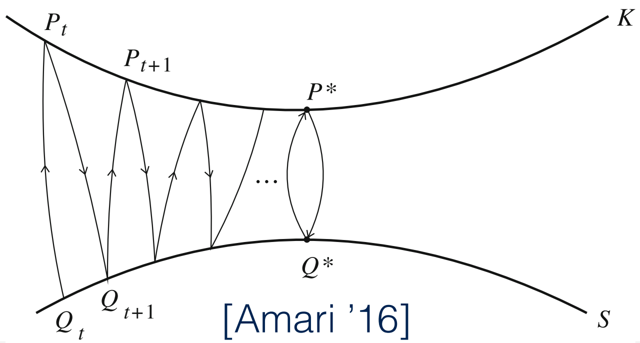
Information Geometric Perspectives

$$\frac{\partial G}{\partial w_{ij}} = -\frac{1}{T}(p_{ij} - p'_{ij})$$

[Ackley, Hinton, Sejnowski '85]



[Amari, Kurata, Nagaoka '92]



[Amari '16]

With hidden units $x = (x_V, x_H)$

$$p_\theta(x_V) = \sum_{x_H} \exp\left(\sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j - \psi(\theta)\right)$$

- The Boltzmann machine defines a curved manifold with singularities
- MLE minimizes KL-divergence from m-flat *data manifold* to the e-flat fully observable Boltzmann manifold
- Iterative optimization using m- and e-projections, EM-algorithm

[Amari, Kurata, Nagaoka '92]

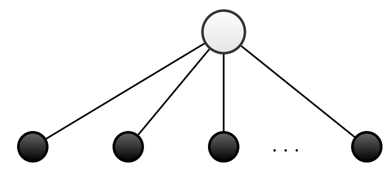
Algebraic Geometric Perspectives

- A Boltzmann machine has a polynomial parametrization and defines a *semialgebraic variety* in the probability simplex
- Main invariant of interest is the *expected dimension* and the number of parameters of (Zariski) dense models
- Implicitization: Find an ideal basis that cuts out the model from the probability simplex

$$\{p = g(\theta) \colon \theta \in \mathbb{R}^d\} \cap \Delta$$

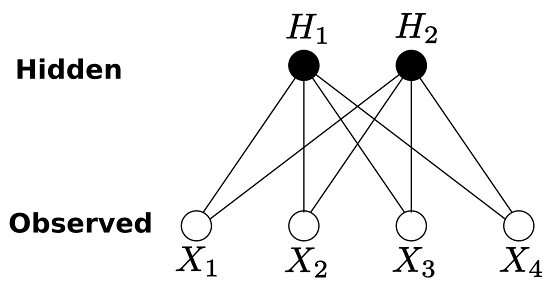
$$\{p \in \Delta \colon f(p) = 0, f \in I\}$$

[Pistone, Riccomagno, Wynn ‘01] [Garcia, Stillman, Sturmfels ‘05]
[Geiger, Meek, Sturmfels ‘06] [Cueto, Morton, Sturmfels ‘10]



3 x 3 minors
of 2-d flattenings

[Raicu ’11]



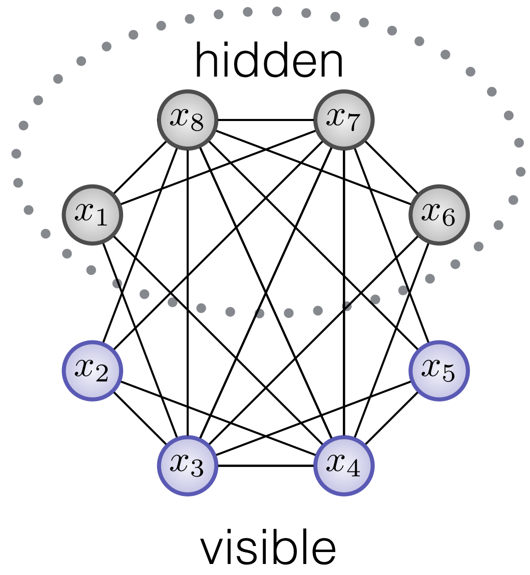
One polynomial of degree 110
and >5.5 trillion monomials

[Cueto, Tobis, Yu ’10]

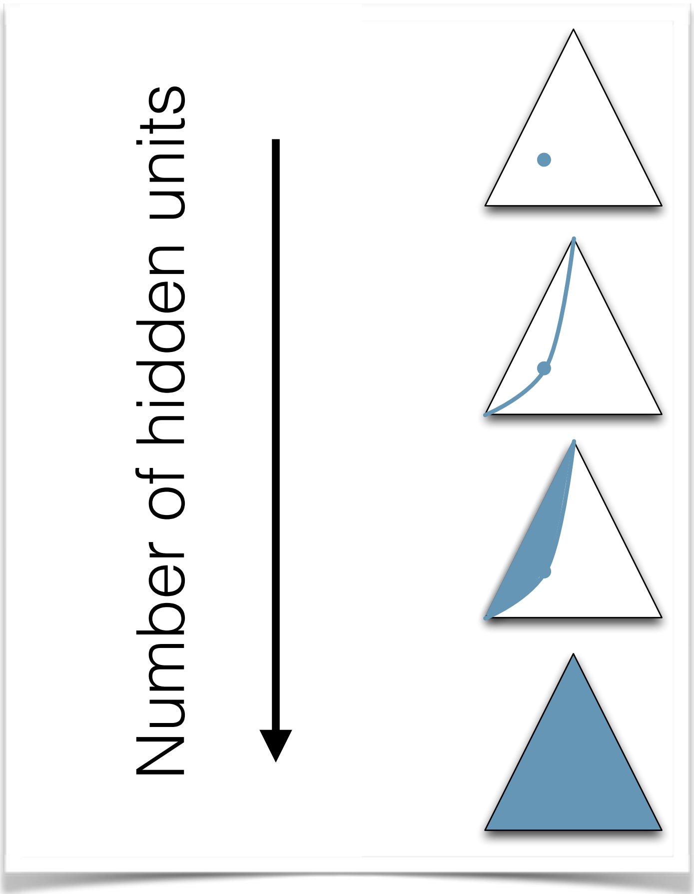
Questions

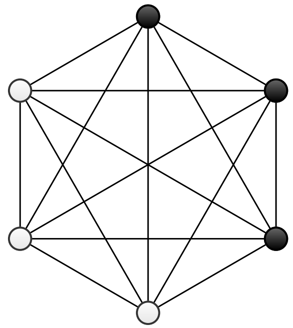
$$p_\theta(x_V) = \sum_{x_H} \exp\left(\sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j - \psi(\theta)\right), \qquad x_V \in \{0,1\}^V$$

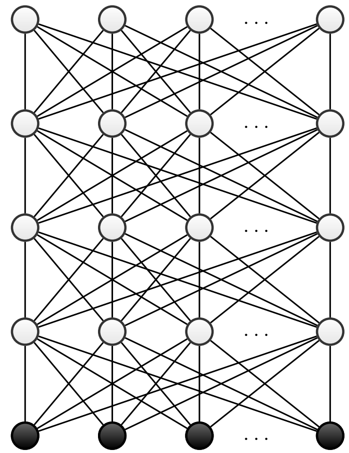
- **Universal Approximation.** What is the smallest number of hidden units such that any distribution on $\{0,1\}^V$ can be represented to within any desired accuracy?
- **Dimension.** What is the dimension of the set of distributions represented by a fixed network?
- **Approximation errors.** MLE, maximum and expected KL-divergence, etc.
- **Support sets.** Properties of the marginal polytopes.

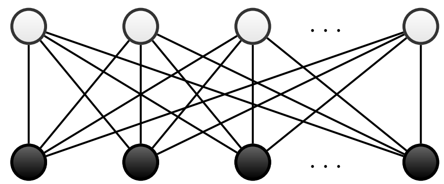


Various Possible Hierarchies

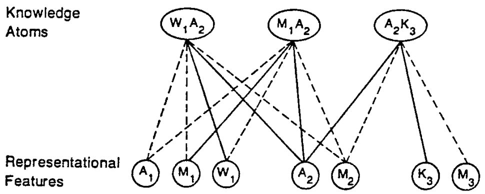



fully connected

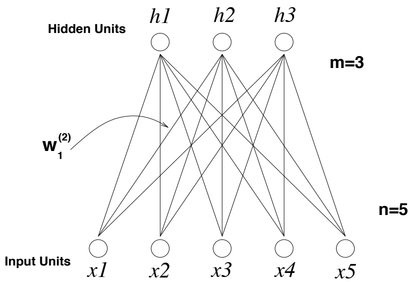

stack of layers


bipartite graph

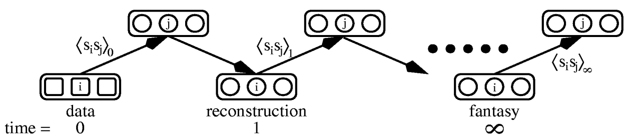
Restricted Boltzmann Machine



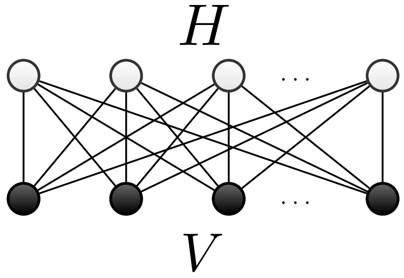
[Smolensky '86]
Harmony Theory



[Freund & Haussler '94]
Influence Combination Machine



[Hinton '02]
Products of Experts



$$\#\text{parameters} = V \cdot H + V + H$$

$$p(x_V | x_H) = \prod_{i \in V} p(x_i | x_H)$$

$$p(x_H | x_V) = \prod_{j \in H} p(x_j | x_V)$$

$$p(x_V) \propto \prod_{j \in H} q_j(x_V)$$

$$q_j(x_V) = \lambda_j \prod_{i \in V} r_{j,i}(x_i) + (1 - \lambda_j) \prod_{i \in V} s_{j,i}(x_i)$$

Universal Approximation

Universal Approximation

Let $H_V := \min\{H : \text{ RBM is a universal approximator on } \{0,1\}^V\}$

nr. parameters behaviour

Observation	$H_V \geq \frac{2^V - V - 1}{V+1}.$	2^V
Theorem (Freund & Haussler '94)	$H_V \leq 2^V.$	$V2^V$
Theorem (Le Roux & Bengio '10)	$H_V \leq 2^V.$	
Theorem (Younes '95)	$H_V \leq 2^V - V - 1.$	
Theorem (M. & Ay '11)	$H_V \leq \frac{1}{2}2^V - 1.$	
Theorem (M. & Rauh '16)	$H_V \leq \frac{2(\log(V)+1)}{V+1}2^V - 1.$	$\log(V)2^V$

Comparison with mixtures of product distributions

Theorem. *Every distribution on* $\{0,1\}^V$ *can be approximated arbitrarily well by a mixture of* k *product distributions if and only if* $k \geq 2^{V-1}$.

$$\Theta(V2^V)$$

[M., Kybernetika '13]

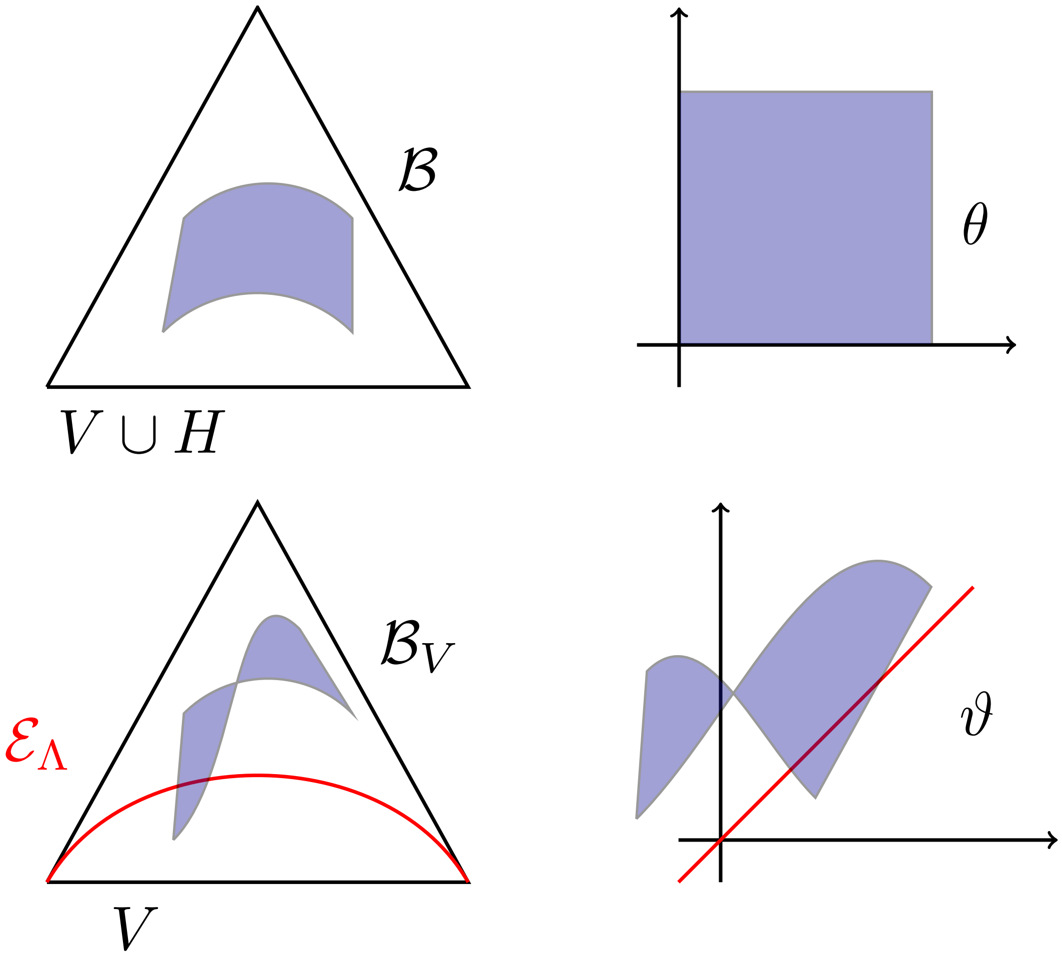
Theorem. *Every distribution on* $\{0,1\}^V$ *can be approximated arbitrarily well by distributions from* $\mathrm{RBM}_{V,H}$ *whenever* $H \geq \frac{2(\log(V-1)+1)}{V+1}(2^V-(V+1)-1)+1$.

$$\Omega(2^V), \quad O(\log(V)2^V)$$

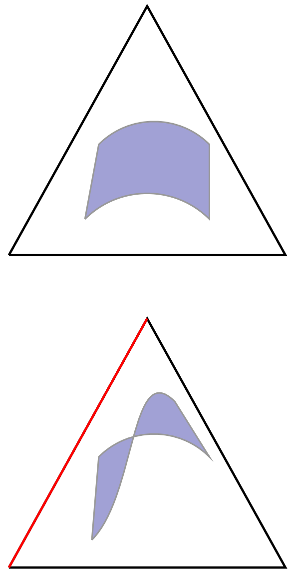
[M. & Rauh '16]

Proof I - Intuition

Each hidden unit extends the RBM along some parameters of the simplex



[M. & Rauh '16]
[Younes '95]



Previous Approach

[M. & Ay '11]
[Le Roux & Bengio '08]

Proof II

Hierarchical models

Consider the set $\mathcal{E}_\Lambda$ of probability vectors

$$q_\vartheta(x_V) = \exp\left(\sum_{\lambda\in\Lambda} \vartheta_\lambda \prod_{i\in\lambda} x_i - \psi(\vartheta)\right), \qquad x_V \in \{0,1\}^V,$$

for all $\vartheta \in \mathbb{R}^\Lambda$, where Λ is an inclusion closed subset of 2^V.

Natural parameters

$$q_\vartheta(x_V) \quad \leftrightarrow \quad -H(x) \quad = \quad \sum_{\lambda\in\Lambda} \vartheta_\lambda \prod_{i\in\lambda} x_i \quad \leftrightarrow \quad (\vartheta_\lambda)_{\lambda\in\Lambda} \in \mathbb{R}^\Lambda, (\vartheta_\lambda)_{\lambda\notin\Lambda} = 0$$

Coordinates for the visible probability simplex

We will use each hidden unit to model a group of monomials

Proof III

Boltzmann Machine

$$p_\theta(x_V) = \sum_{x_H} \exp\Big(\sum_i \theta_i x_i + \sum_{i\in V, j\in H} \theta_{ij} x_i x_j - \psi(\theta)\Big), \quad x_V \in \{0,1\}^V$$

Free Energy

$$\begin{aligned} p_\theta(x_V) \quad \leftrightarrow \quad -F(x_V) &= \log\left(\sum_{x_H} \exp\Big(\sum_i \theta_i x_i + \sum_{i\in V, j\in H} \theta_{ij} x_i x_j\Big)\right) \\ &= \sum_{j\in H} \log\Big(1 + \exp(\theta_j + \sum_{i\in V} \theta_{ij} x_i)\Big) \end{aligned}$$

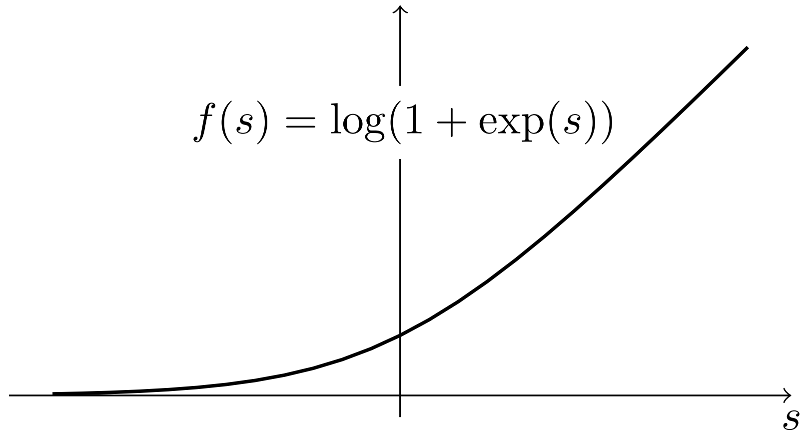
Natural parameters in the visible probability simplex

$$\leftrightarrow \quad \vartheta_B(\theta) = \sum_{j\in H} \sum_{C\subseteq B} (-1)^{|B\setminus C|} \log\Big(1 + \exp(\theta_j + \sum_{i\in C} \theta_{ij})\Big), \quad B \in 2^V$$

Sum of independent terms

Proof IV - Softplus polynomials

$$\varphi(x_V) = \log\Big(1 + \exp(\theta_j + \sum_{i \in V} \theta_{ij} x_i)\Big)$$
$$= \sum_{B \subseteq V} K_{j,B} \prod_{i \in B} x_i$$

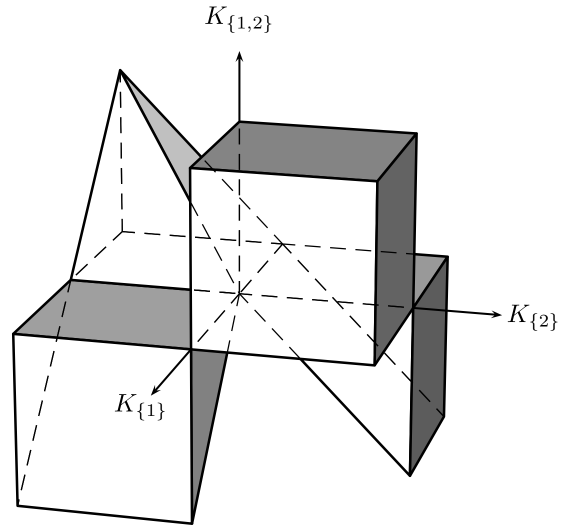


We show that certain groups of coefficients can be made arbitrary:

Lemma 2. *Consider an edge pair (B, B'). Depending on $|B|$, for any $\epsilon > 0$ there is a choice of $w_B \in \mathbb{R}^B$ and $c \in \mathbb{R}$ such that $\|(K_B, K_{B'}) - (J_B, J_{B'})\| \leq \epsilon$ if and only if*

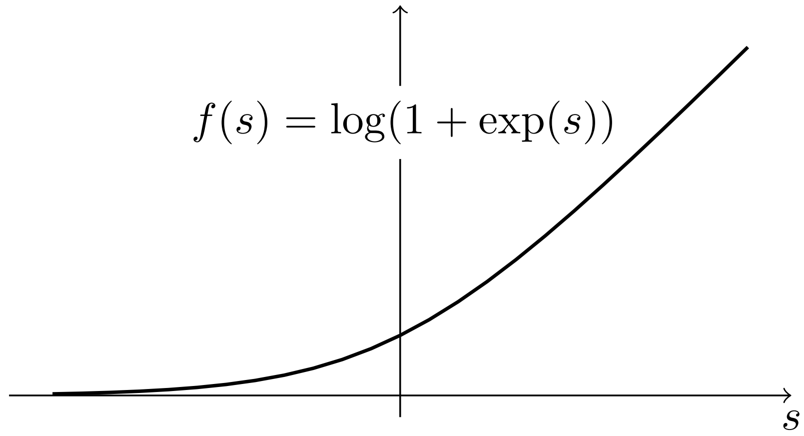
$$\begin{aligned}
&J_{B'} \geq 0, -J_B, && \text{for } |B| = 1\\
&J_{B'} \geq 0, -J_B \quad \text{or} \quad J_{B'} \leq 0, -J_B, && \text{for } |B| = 2\\
&J_{B'} \geq 0, -J_B \quad \text{or} \quad J_{B'} \leq 0, -J_B, && \text{for } |B| = 3\\
&(J_B, J_{B'}) \in \mathbb{R}^2, && \text{for } |B| \geq 4.
\end{aligned}$$

Lemma 5. *Consider any $B, B' \subseteq V$ with $B \cap B' = \emptyset$. Let $w_i = 0$ for $i \notin B \cup B'$. Then, for any $J_{B\cup\{j\}} \in \mathbb{R}$, $j \in B'$, and $\epsilon > 0$, there is a choice of $w_{B\cup B'} \in \mathbb{R}^{B\cup B'}$ and $c \in \mathbb{R}$ such that $|K_{B\cup\{j\}} - J_{B\cup\{j\}}| \leq \epsilon$ for all $j \in B'$, and $|K_C| \leq \epsilon$ for all $C \neq B, B \cup \{j\}$, $j \in B'$.*



Proof IV - Softplus polynomials

$$\begin{aligned}\varphi(x_V) &= \log\Big(1+\exp(\theta_j + \sum_{i\in V}\theta_{ij}x_i)\Big)\\ &= \sum_{B\subseteq V} K_{j,B}\prod_{i\in B} x_i\end{aligned}$$

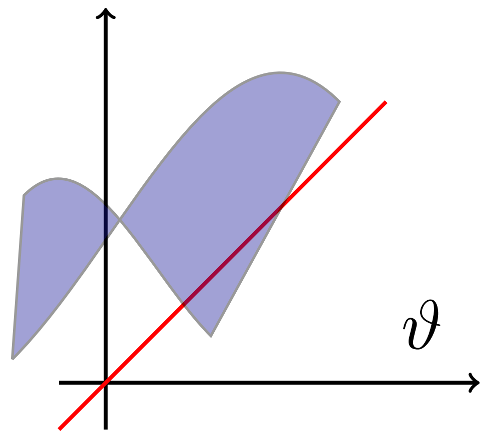


We show that certain groups of coefficients can be made arbitrary:

Lemma 2. *Consider an edge pair (B, B'). Depending on $|B|$, for any $\epsilon > 0$ there is a choice of $w_B \in \mathbb{R}^B$ and $c \in \mathbb{R}$ such that $\|(K_B, K_{B'}) - (J_B, J_{B'})\| \leq \epsilon$ if and only if*

$$\begin{array}{lll} J_{B'} \geq 0, -J_B, & & \text{for } |B| = 1\\ J_{B'} \geq 0, -J_B & \text{or} \quad J_{B'} \leq 0, -J_B, & \text{for } |B| = 2\\ J_{B'} \geq 0, -J_B & \text{or} \quad J_{B'} \leq 0, -J_B, & \text{for } |B| = 3\\ (J_B, J_{B'}) \in \mathbb{R}^2, & & \text{for } |B| \geq 4.\end{array}$$

Lemma 5. *Consider any $B, B' \subseteq V$ with $B \cap B' = \emptyset$. Let $w_i = 0$ for $i \notin B \cup B'$. Then, for any $J_{B\cup\{j\}} \in \mathbb{R}$, $j \in B'$, and $\epsilon > 0$, there is a choice of $w_{B\cup B'} \in \mathbb{R}^{B\cup B'}$ and $c \in \mathbb{R}$ such that $|K_{B\cup\{j\}} - J_{B\cup\{j\}}| \leq \epsilon$ for all $j \in B'$, and $|K_C| \leq \epsilon$ for all $C \neq B, B\cup\{j\}$, $j \in B'$.*

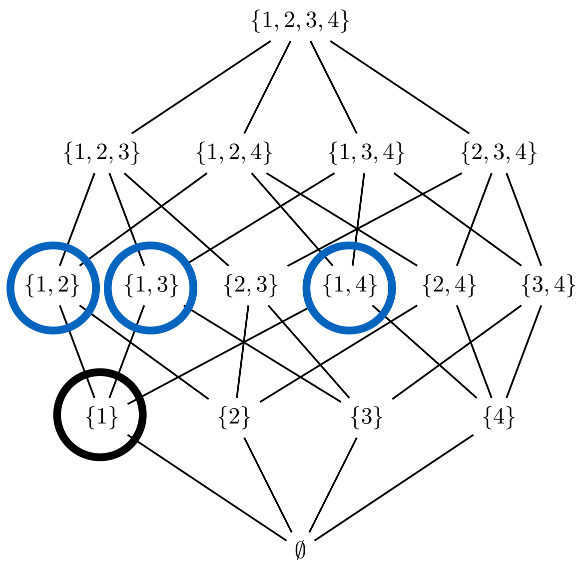


Proof V - Coverings

- Each hidden unit adds a linear space of coefficients, corresponding to an exponential family of dim up to V

- Adding sufficiently many linear spaces produces any hierarchical model

- Previous proofs added at most 1 or 2 dimensions per hidden unit



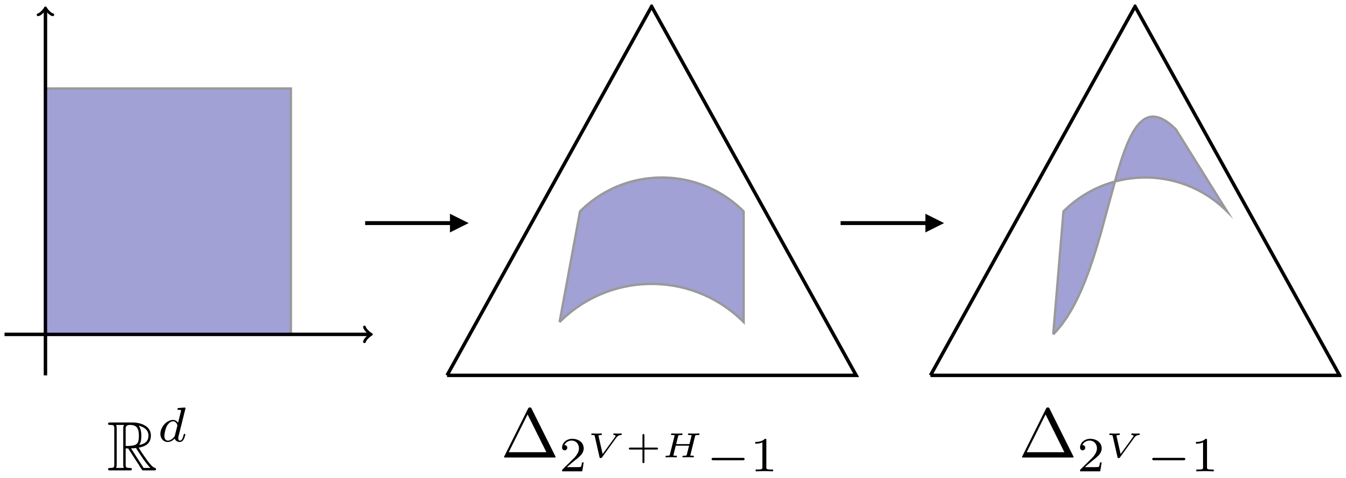
Theorem. *Let* $1 \leq k \leq V$. *Every distribution from the* k*-interaction model* $\mathcal{E}_k$ *on* $\{0,1\}^V$ *can be approximated arbitrarily well by distributions from* $\mathrm{RBM}_{V,H}$ *whenever* $H \geq \frac{\log(V-1)+1}{V+1} \sum_{s=2}^{k} \binom{V+1}{s}$.

QED

Dimension

Dimension

Consider $\mathcal{M} = \{p_\theta : \theta \in \mathbb{R}^d\} \subseteq \Delta_{N-1}$ parametrized by $\phi : \mathbb{R}^d \to \Delta_{N-1};\ \theta \mapsto p_\theta$.



Conjecture (Cueto, Morton, Sturmfels, 2010)**.** *The restricted Boltzmann machine has the expected dimension, i.e., it is a semialgebraic set of dimension* $\min\{VH + V + H, 2^V - 1\}$ *in* Δ_{2^V-1}.

Dimension

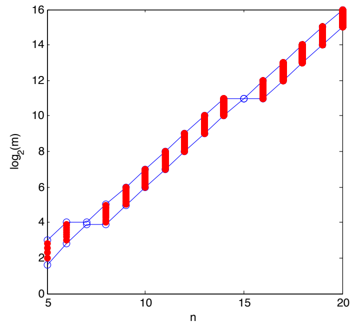
Theorem (Cueto, Morton, Sturmfels, 2010). *The restricted Boltzmann machine has the expected dimension* $\min\{VH+V+H, 2^V-1\}$ *when* $H \le 2^{V-\lceil \log_2(V+1)\rceil}$ *and when* $H \ge 2^{V-\lfloor \log_2(V+1)\rfloor}$.

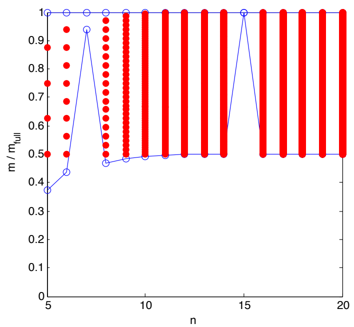
Special Cases

n	$k \le$	$k \ge$
5	2^{2}	$\mathbf{7}$
6	2^{3}	$\mathbf{12}$
7	2^{4}	2^{4}
8	$\mathbf{2^{2}\cdot 5}$	2^{5}
9	$\mathbf{2^{3}\cdot 5}$	$\mathbf{62}$
10	$\mathbf{2^{3}\cdot 9}$	$\mathbf{120}$
11	$\mathbf{2^{4}\cdot 9}$	$\mathbf{192}$
12	2^{8}	$\mathbf{380}$
13	2^{9}	$\mathbf{736}$
14	2^{10}	$\mathbf{1408}$
15	2^{11}	2^{11}
16	$\mathbf{2^{5}\cdot 85}$	2^{12}
17	$\mathbf{2^{6}\cdot 83}$	2^{13}
18	$\mathbf{2^{8}\cdot 41}$	2^{14}
19	$\mathbf{2^{12}\cdot 5}$	$\mathbf{31744}$
20	$\mathbf{2^{12}\cdot 9}$	$\mathbf{63488}$
21	$\mathbf{2^{13}\cdot 9}$	$\mathbf{122880}$
22	$\mathbf{2^{14}\cdot 9}$	$\mathbf{245760}$
23	$\mathbf{2^{15}\cdot 9}$	$\mathbf{393216}$
24	2^{19}	$\mathbf{786432}$
25	2^{20}	$\mathbf{1556480}$
26	2^{21}	$\mathbf{3112960}$
27	2^{22}	$\mathbf{6029312}$
28	2^{23}	$\mathbf{12058624}$
29	2^{24}	$\mathbf{23068672}$
30	2^{25}	$\mathbf{46137344}$
31	2^{26}	2^{26}
32	$\mathbf{2^{20}\cdot 85}$	2^{27}
33	$\mathbf{2^{21}\cdot 85}$	2^{28}

n	$k \le$
35	$\mathbf{2^{23}\cdot 83}$
37	$\mathbf{2^{26}\cdot 41}$
39	$\mathbf{2^{31}\cdot 5}$
47	$\mathbf{2^{38}\cdot 9}$
63	2^{57}
70	$\mathbf{2^{43}\cdot 1657009}$
71	$\mathbf{2^{63}\cdot 3}$
75	$\mathbf{2^{63}\cdot 41}$
79	$\mathbf{2^{70}\cdot 5}$
95	$\mathbf{2^{85}\cdot 9}$
127	2^{120}
141	$\mathbf{2^{113}\cdot 1657009}$
143	$\mathbf{2^{134}\cdot 3}$
151	$\mathbf{2^{138}\cdot 41}$
159	$\mathbf{2^{149}\cdot 5}$
163	$\mathbf{2^{151}\cdot 19}$
191	$\mathbf{2^{180}\cdot 9}$
255	2^{247}
270	$\mathbf{2^{202}\cdot 1021273028302258913}$
283	$\mathbf{2^{254}\cdot 1657009}$
287	$\mathbf{2^{277}\cdot 3}$
300	$\mathbf{2^{220}\cdot 3348824985082075276195}$
303	$\mathbf{2^{289}\cdot 41}$
319	$\mathbf{2^{308}\cdot 5}$
327	$\mathbf{2^{314}\cdot 19}$
383	$\mathbf{2^{371}\cdot 9}$
511	2^{502}
512	$\mathbf{2^{443}\cdot 1021273028302258913}$

Open Cases





Theorem (M. & Morton, 2016). *The restricted Boltzmann machine has the expected dimension* $\min\{VH+V+H, 2^V-1\}$.

Proof I - Marginals of Exponential Families

Let $\mathcal{M}_F$ be given by

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

Dimension is maximum rank of Jacobian matrix

$$J_{\mathcal{M}_F}(\theta) = \left(\sum_y p_\theta(x,y) F(x,y) - \sum_y p_\theta(x,y) \sum_{x',y'} p_\theta(x',y') F(x',y') \right)_x$$

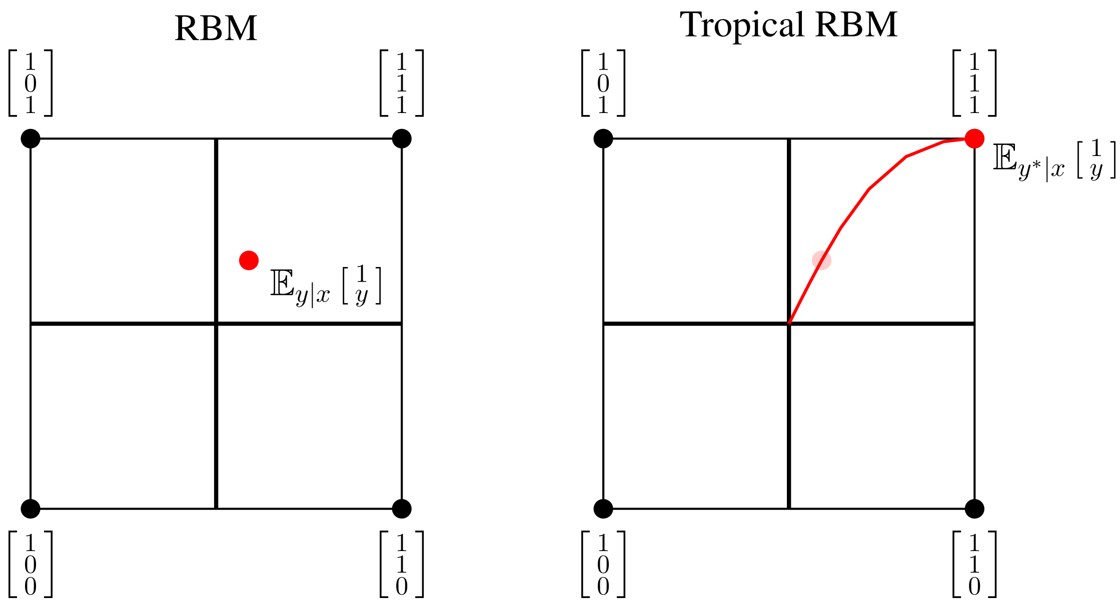
$$\begin{aligned} \operatorname{rank}\left(J_{\mathcal{M}_F}(\theta)\right) &= \operatorname{rank} \left(\sum_y p_\theta(x,y) F(x,y) \right)_x - 1 \\ &= \operatorname{rank} \left(\sum_y p_\theta(y|x) F(x,y) \right)_x - 1. \end{aligned}$$

expectation parameters of conditional distributions

Tropical Dimension Approach - Intuitive View

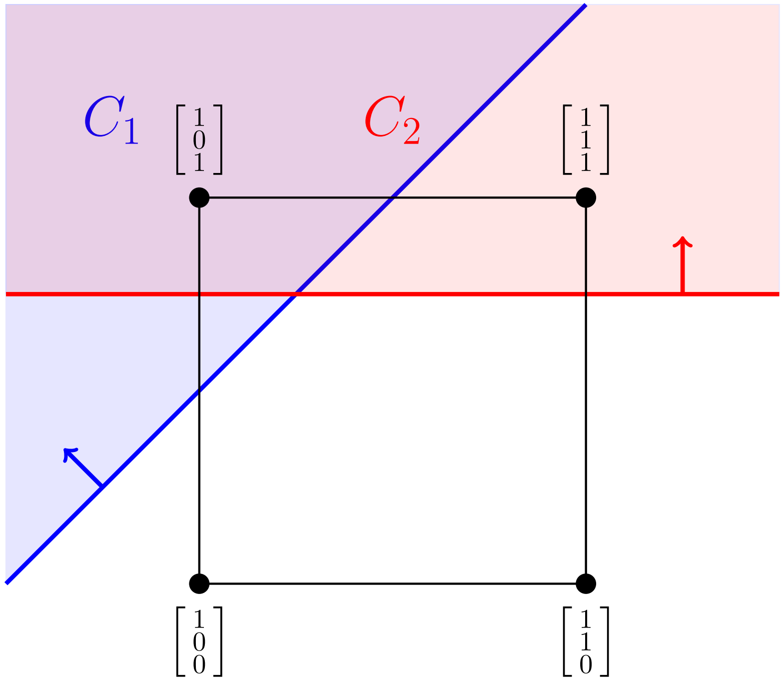
$$\max_{\theta} \ \operatorname{rank} \left(\sum_{y} p_{\theta}(x|y) F(x,y) \right)_{x} \geq \max_{\theta} \ \operatorname{rank} \left(F(x, h_{\theta}(x)) \right)_{x}$$

$$h_{\theta}(x) := \operatorname{argmax}_{y} p_{\theta}(y|x) = \operatorname{argmax}_{y} \langle \theta, F(x,y) \rangle$$



[Bieri-Groves ‘84] [Draisma ‘08] [Cueto, Morton, Sturmfels ‘10] [M. & Morton ’15]

Tropical Dimension Approach - Intuitive View



$$J_{\mathrm{RBM}_{n,m}^{\mathrm{tropical}}}(W, b, c) = \left[\begin{array}{c} X \\ \hline X_{C_1} \\ \hline \vdots \\ \hline X_{C_m} \end{array}\right]$$

- Tropical approach is very powerful. In many cases the tropical rank is associated to known combinatorial quantities
- However, many cases it leads to very hard combinatorial problems

Proof II

Theorem (Catalisano, Geramita, Gimigliano, 2011 - rephrased). *The set of mixtures of* $H+1$ *product distributions of* V *binary variables has the expected dimension* $\min\{VH+V+H, 2^V-1\}$*, whenever* $V \geq 5$.

Observation. *The sufficient statistics matrix of* $\mathrm{RBM}_{V,H}$ *satisfies* $F(x,y) = A(x) \otimes B(y)$*, where* A, B *describe* V *and* H *independent binary variables and each includes a constant row.*

Lemma. *Let* A, B, C *be sufficient statistics matrices, each containing a constant row. If* B *describes* H *independent binary variables and* C *describes one categorical variable with* $H+1$ *values, then* $\dim(\mathcal{M}_{A\otimes B}) \geq \dim(\mathcal{M}_{A\otimes C})$.

Proof III

- For the RBM we have

$$\operatorname{rank}\left(J_{\mathrm{RBM}_{n,m}}(\theta)\right) = \operatorname{rank}\left(\begin{bmatrix}1\\x\end{bmatrix} \otimes \mathbb{E}_{y|x}\begin{bmatrix}1\\y\end{bmatrix}\right)_x .$$

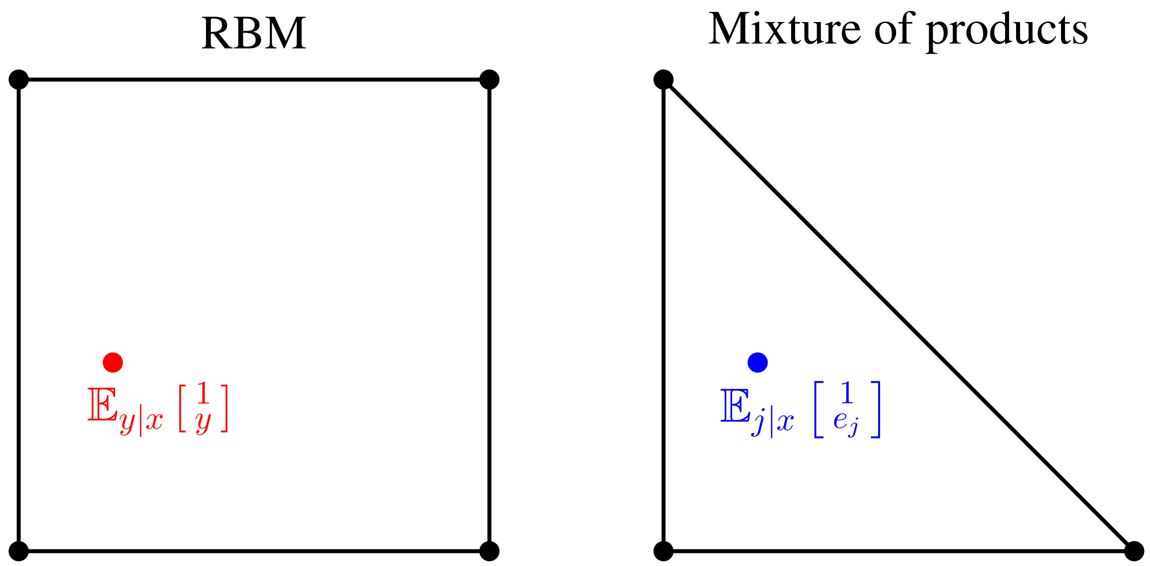
- For the mixture of products we have

$$\operatorname{rank}\left(J_{\mathrm{M}_{n,m+1}}(\theta)\right) = \operatorname{rank}\left(\begin{bmatrix}1\\x\end{bmatrix} \otimes \mathbb{E}_{j|x}\begin{bmatrix}1\\e_j\end{bmatrix}\right)_x .$$

- We show that to any $J_{\mathrm{Mixt}_{n,m+1}}(\theta)$ there is a $J_{\mathrm{RBM}_{n,m}}(\theta)$ with the same rank.

Proof IV

$$\mathbb{E}_{y|x}\begin{bmatrix}1\\y\end{bmatrix} = \begin{bmatrix}1\\p_\theta(y_1 = 1|x)\\\vdots\\p_\theta(y_m = 1|x)\end{bmatrix} \qquad \mathbb{E}_{j|x}\begin{bmatrix}1\\e_j\end{bmatrix} = \begin{bmatrix}1\\\tilde{p}_\theta(1|x)\\\vdots\\\tilde{p}_\theta(m|x)\end{bmatrix}$$



QED

Conclusion

- Boltzmann machines define marginals of exponential families with an interesting geometry.

- I presented new results on two basic questions:

  Universal approximation
  RBMs and BMs are universal approximators with significantly less parameters than previously known.
  This result also shows that universal approximation with RBMs require significantly less parameters than with mixtures of products

  Dimension
  RBMs always have the expected dimension.
  This completes the dimension characterization initiated by Cueto, Morton, Sturmfels, and resolves their conjecture positively

Open Problems

- Can the universal approximation bounds for restricted Boltzmann machines be improved?
- Do deep Boltzmann machines have the expected dimension?
- Are less parameters possible with deep Boltzmann machines?

Literature

Literature

Montúfar & Rauh, *Hierarchical Models as Marginals of Hierarchical Models*, **arXiv:1508.03606v2**

Montúfar & Morton, *Dimension of Marginals of Kronecker Product Models,* **arXiv:1511.03570**

Related Literature

Amari, Kurata, Nagaoka, Information Geometry of Boltzmann Machines, IEEE Transactions on Neural Networks, Vol 3 Issue 2, pp. 260-271,1992

Younes, *Synchronous Boltzmann Machines can be Universal Approximators*, Applied Mathematics Letters 9: 109-113,1996

Cueto, Morton, Sturmfels, *Geometry of the Restricted Boltzmann Machine*, Algebraic Methods in Statistics and Probability 2, AMS Special Session, 2010

Montúfar, *Universal approximation depth and errors of narrow belief networks with discrete units,* Neural Computation 26: 1386-1407, 2014

Montúfar & Ay, *Refinements of universal approximation results for DBNs and RBMs,* Neural Computation 23: 1306-1319, 2011

Montúfar, *Mixture decompositions of exponential families using a decomposition of their sample spaces,* Kybernetika 49: 23-39, 2013

Montúfar & Morton, *Discrete Restricted Boltzmann Machines,* JMLR 16: 653-672, 2015