An Algebraic Perspective on Deep Learning

Jason Morton

Penn State

July 19-20, 2012
IPAM

Supported by DARPA FA8650-11-1-7145.

Motivating Questions

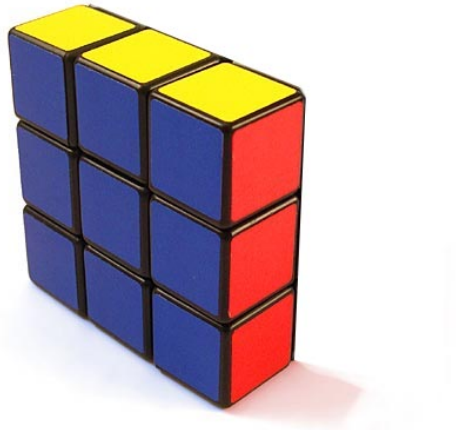
- Characterize representational power and learning performance.
  - What can DL models model? How well?
  - Can we bound approximation errors, prove convergence, etc.?
- Seek *a priori* design insights
  - What model architecture for what data?
  - Can we predict performance?
  - What tradeoffs should we make?
- Understand representations obtained
  - Identifiability
  - Transferrability
- I'll describe an algebraic approach to these kinds of problems in three parts

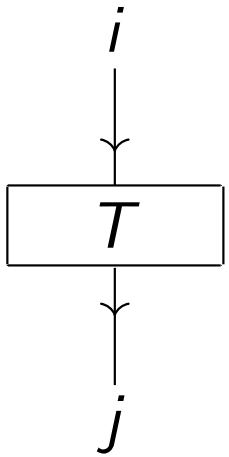
Outline

1. Algebraic geometry of tensor networks
   - Tensors
   - Tensor Networks
   - Algebraic geometry
2. Algebraic Description of Graphical Models
   - Review of GM Defininitions
   - Algebraic and semialgebraic descriptions
   - Restricted Boltzmann machines
3. Identifiability, singular learning theory, other perspectives
   - Identifiability
   - Singular Learning Theory

1. Algebraic geometry of tensor networks
   - Tensors
   - Tensor Networks
   - Algebraic geometry
2. Algebraic Description of Graphical Models
   - Review of GM Defininitions
   - Algebraic and semialgebraic descriptions
   - Restricted Boltzmann machines
3. Identifiability, singular learning theory, other perspectives
   - Identifiability
   - Singular Learning Theory

Matrices





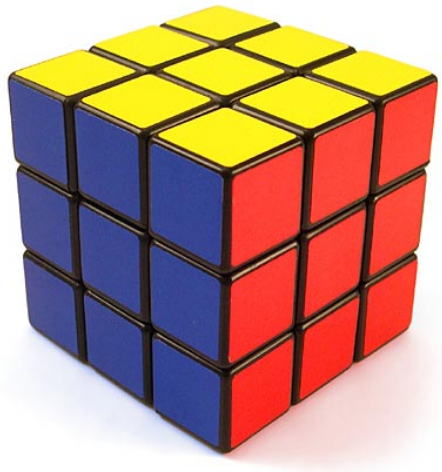
A matrix $M = (m_{ij})$

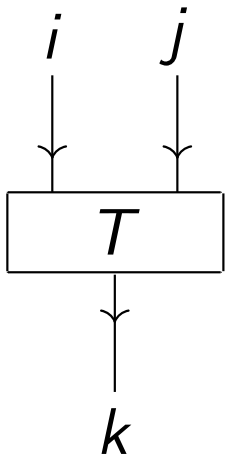
- represents a linear transformation $\mathbf{U} \to V$
- is a 2-way array
- has an action by $\mathrm{GL}(\mathbf{U})$, $\mathrm{GL}(V)$ on two sides

Matrix decomposition is a workhorse of ML, much else

Most data can be flattened into matrix format

Tensors





A tensor $T = (t_{ijk})$

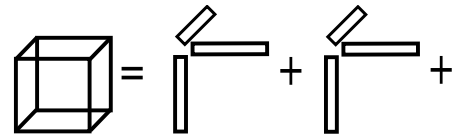
- represents a multilinear transformation $U \otimes V \to W$, $W \to U \otimes V$, $U \otimes W \to V$, etc.
- is a multi-way array (here 3-way)
- with a multilinear action on each "leg" or "side"

Tensor decomposition is possible but more subtle

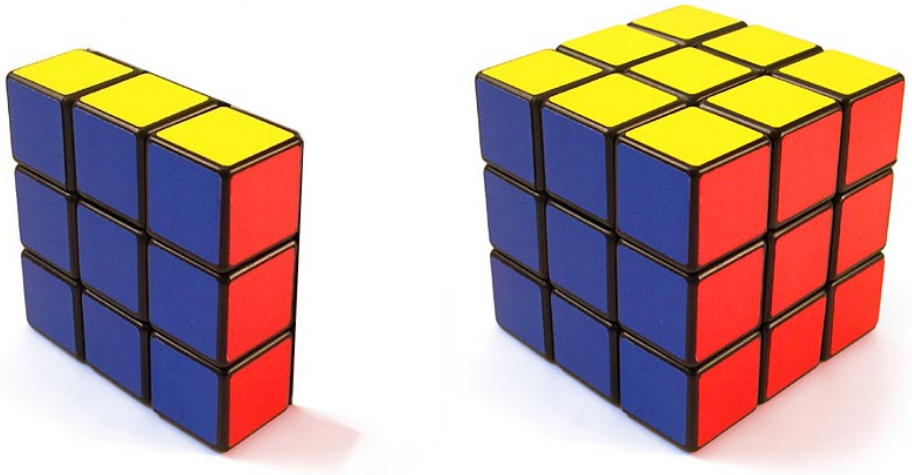
Data arrives in tensor format, and something is lost by flattening

Example of tensor decomposition

Three possible generalizations of eigenvalue decomposition are the same in the matrix case but not in the tensor case. For a $p \times p \times p$ tensor K,

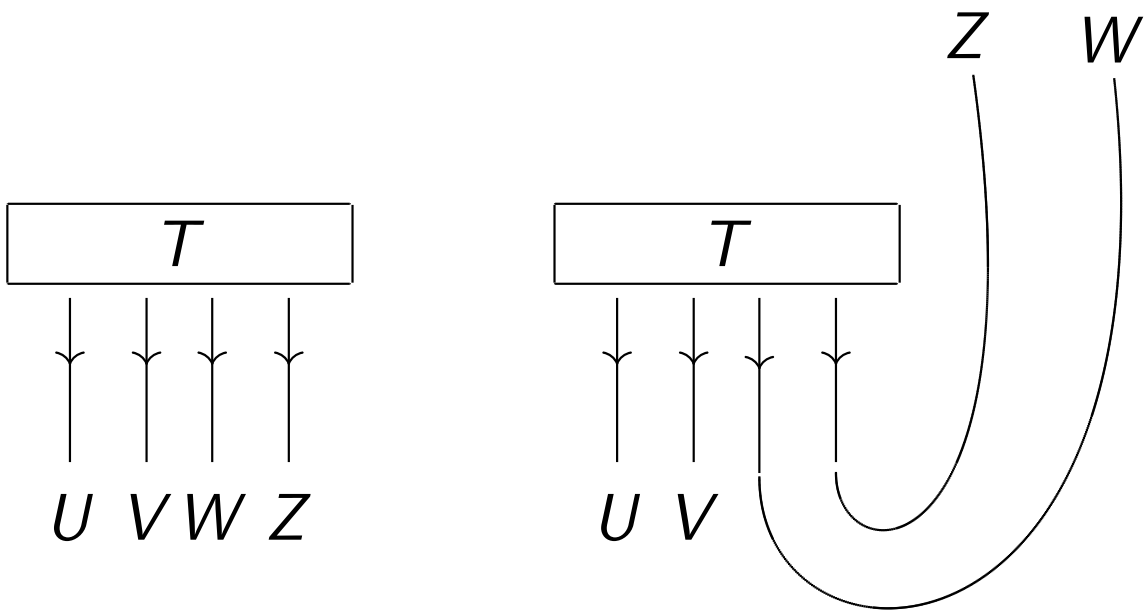
Name	minimum r such that
Tensor rank	$K = \sum_{i=1}^{r} u_i \otimes v_i \otimes w_i$ not closed 
Border rank	$K = \lim_{\epsilon \to 0}(S_\epsilon)$, Tensor rank$(S_\epsilon) = r$ closed but hard to represent; defining equations unknown.
Multilinear rank	$K = A \cdot C, C \in \mathbb{R}^{r \times r \times r}, A \in \mathbb{R}^{p \times r}$, closed and understood.

Matrices vs Tensors



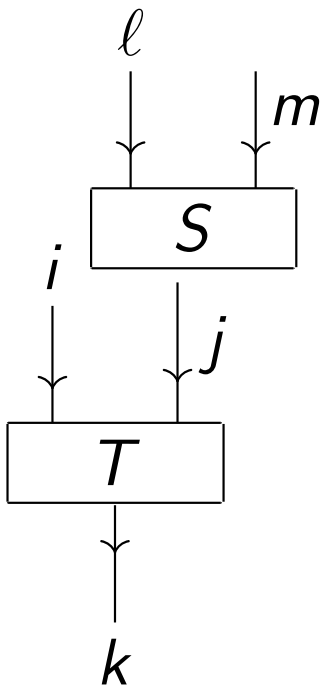
- Generalization of matrix concepts to tensors is usually not straightforward, but
- flattenings are still matrices
- effective computations in multilinear algebra generally reduce to linear algebra (so far)

Flatten



View $T \in U \otimes V \otimes W \otimes Z$ as $T : Z^* \otimes W^* \to U \otimes V$

Contract



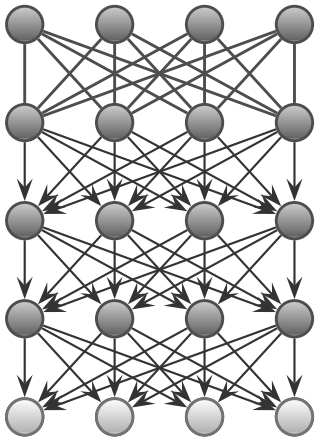
Can express as: flatten, then multiply the matrices, then reshape

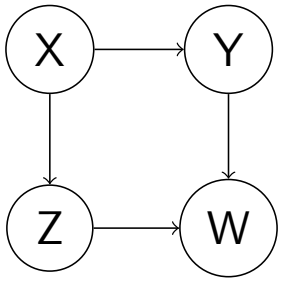
Algebraic geometry of ~~tensor~~ networks

1. Algebraic geometry of tensor networks
   - Tensors
   - Tensor Networks
   - Algebraic geometry
2. Algebraic Description of Graphical Models
   - Review of GM Defininitions
   - Algebraic and semialgebraic descriptions
   - Restricted Boltzmann machines
3. Identifiability, singular learning theory, other perspectives
   - Identifiability
   - Singular Learning Theory

What is a tensor network?

And why do they keep coming up?



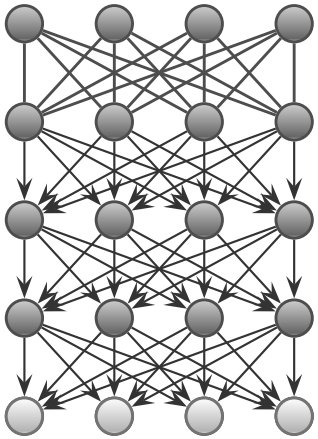


- When we reason about processes in space and time, involving interaction and independence, causality and locality, we tend to draw diagrams (networks) involving boxes, wires, arrows
- Attaching mathematical meaning to such a diagram
  - allows us to quantitatively model real systems and to compute
  - usually leads to defining a monoidal category
- Analyzing the monoidal category often means defining a functor to the category of vector spaces and linear transformations

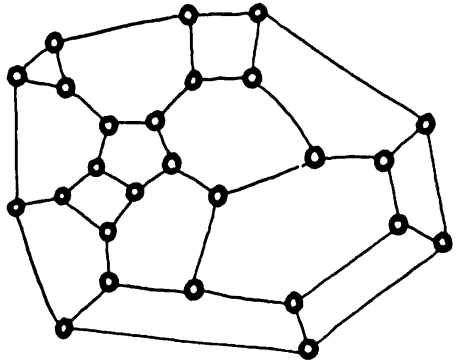
Tensor Networks

- Category theory provides a succint and beautiful means to formalize ideas about such diagrams and their interpetations in applied mathematics
- But I'll focus on diagrams in the category of vector spaces and linear transformations, often called tensor networks
- Fortunately, many of the same mathematical ideas work to analyze these whether they occur in machine learning, statistics, computational complexity, or quantum information
- Algebraic geometry and representation theory provide a powerful set of tools to characterize and understand these objects as they arise in ...

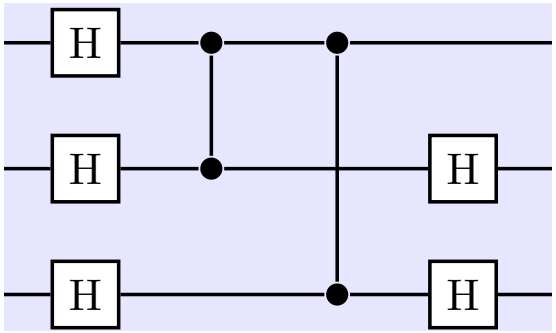
Tensor Networks



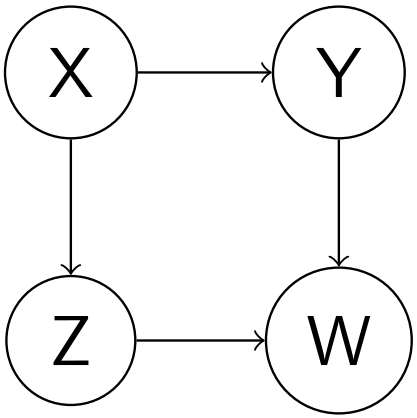
ML and Statistics

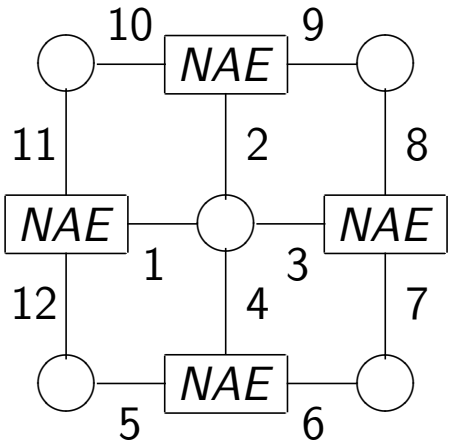


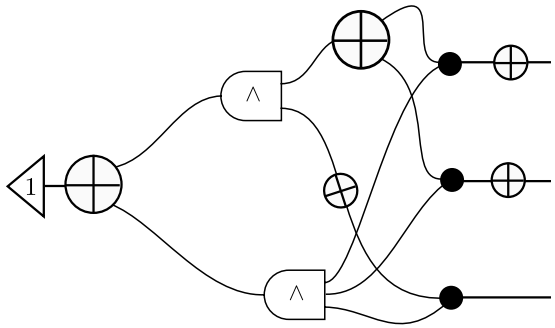
Complexity Theory



Quantum Information and Many-Body Systems

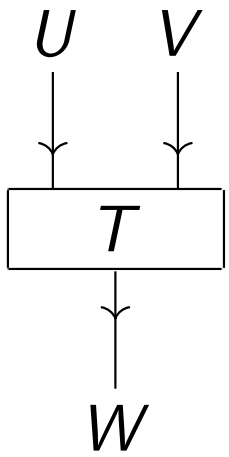






Tensors and wiring diagrams

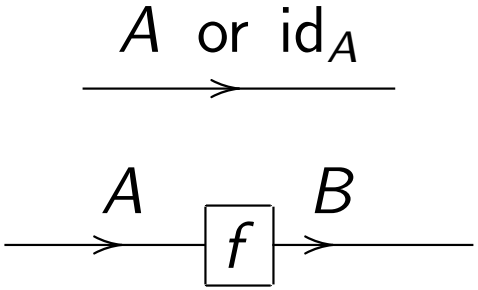
A multilinear operator
$T : U \otimes V \to W$
is a tensor

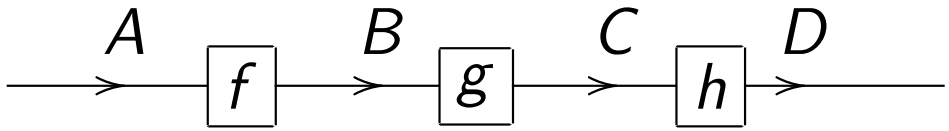


- Draw a wire for each vector space (variable)
- Box for each tensor (factor)
- Arrows denote primal/dual

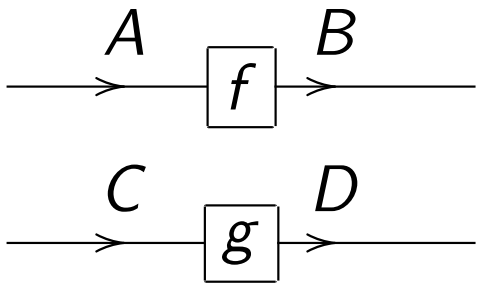
Composition

- Connect wires to compose/contract: $h \circ g \circ f$





- juxtpose to tensor/run in parallel $f \otimes g$



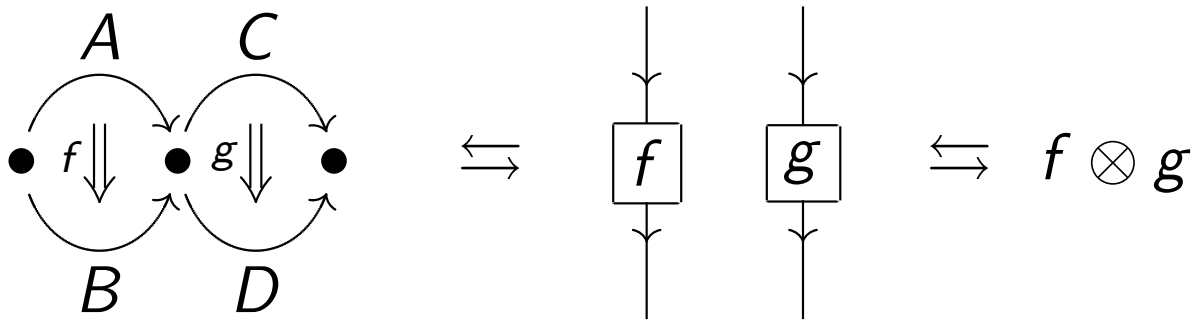
From these primitives (and duals, swaps, special maps) can build complex networks

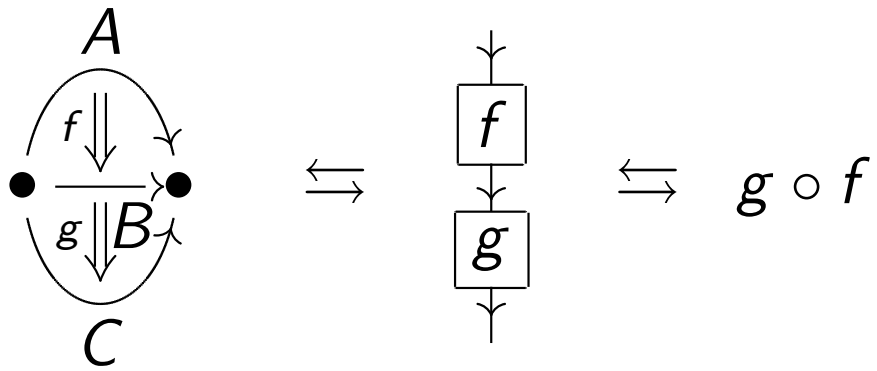
You could have invented graphical models

- This graphical notation has arisen several times in several areas (Feynman diagrams, graphical models, circuits, representation theory...)
- This convergent evolution is no accident and reflects an underlying mathematical structure (monoidal categories)
- Using the resulting common generalizations to study graphical models is promising; makes translating results from/to other areas much easier

Graphical language for monoidal categories

A 2-category with one object is a strict monoidal category; the graphical language is Poincaré dual to the 2-cell diagram notation.





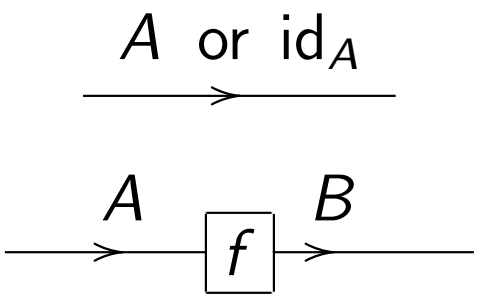
See papers by Joyal and Street, Selinger.

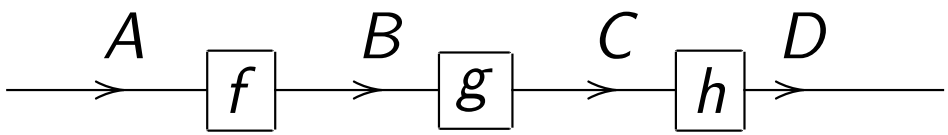
Categories

A category $\mathcal{C}$ consists of a

- class of objects $\mathrm{Ob}(\mathcal{C})$ and set $\mathrm{Mor}(A, B)$ of morphisms for each ordered pair of objects,
- an associative composition rule taking morphisms $A \xrightarrow{f} B$, $B \xrightarrow{g} C$, to a morphism $g \circ f : A \to C$
- an identity $\mathrm{id}_A \in \mathrm{Mor}(A, A)$ such that $\mathrm{id}_B \circ f = f = f \circ \mathrm{id}_A$.

Diagrammatically:





Think of categories and variations as concrete and combinatorial.

Monoidal categories

Definition

A monoidal category $(\mathcal{C}, \otimes, \alpha, \lambda, \rho, \mathrm{I})$ is a category $\mathcal{C}$ with the additional data of

(i) an abstract tensor product functor $\otimes : \mathcal{C} \times \mathcal{C} \to \mathcal{C}$,

(ii) a natural isomorphism called the associator $\alpha_{ABC} : (A \otimes B) \otimes C \to A \otimes (B \otimes C)$,

(iii) a unit object I and natural isomorphisms $\lambda_A : \mathrm{I} \otimes A \to A$ and $\rho_A : A \otimes \mathrm{I}$, the left and right unitors,

such that if w, w' are two words obtained from $A_1 \otimes A_2 \otimes \cdots A_n$ by inserting Is and balanced parenthesis, then all isomorphisms $\phi : w \to w'$ composed of αs, λs, and ρs and their inverses are equal. Thus we have a unique natural transformation $w \to w'$.

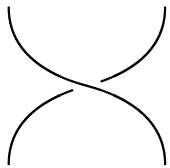
A monoidal category is strict if α, λ, and ρ are equalities. Monoidal categories can be "strictified," so the λ, ρ, α can often be ignored.

Monoidal categories ctd.

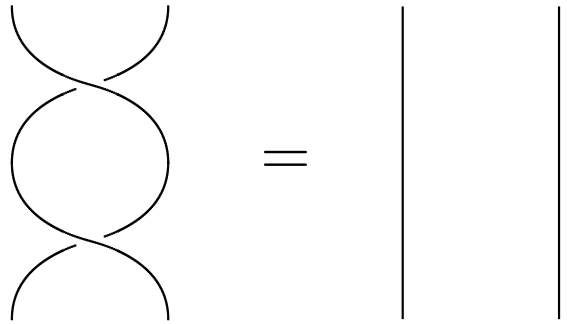
Definition

A monoidal category is braided if it is equipped with natural isomorphisms $b_{A\otimes B} : A \otimes B \xrightarrow{\sim} B \otimes A$ subject to the hexagon axioms and symmetric if $b_{A\otimes B} b_{B\otimes A} = \mathrm{id}_{A\otimes B}$.

Diagramatically:



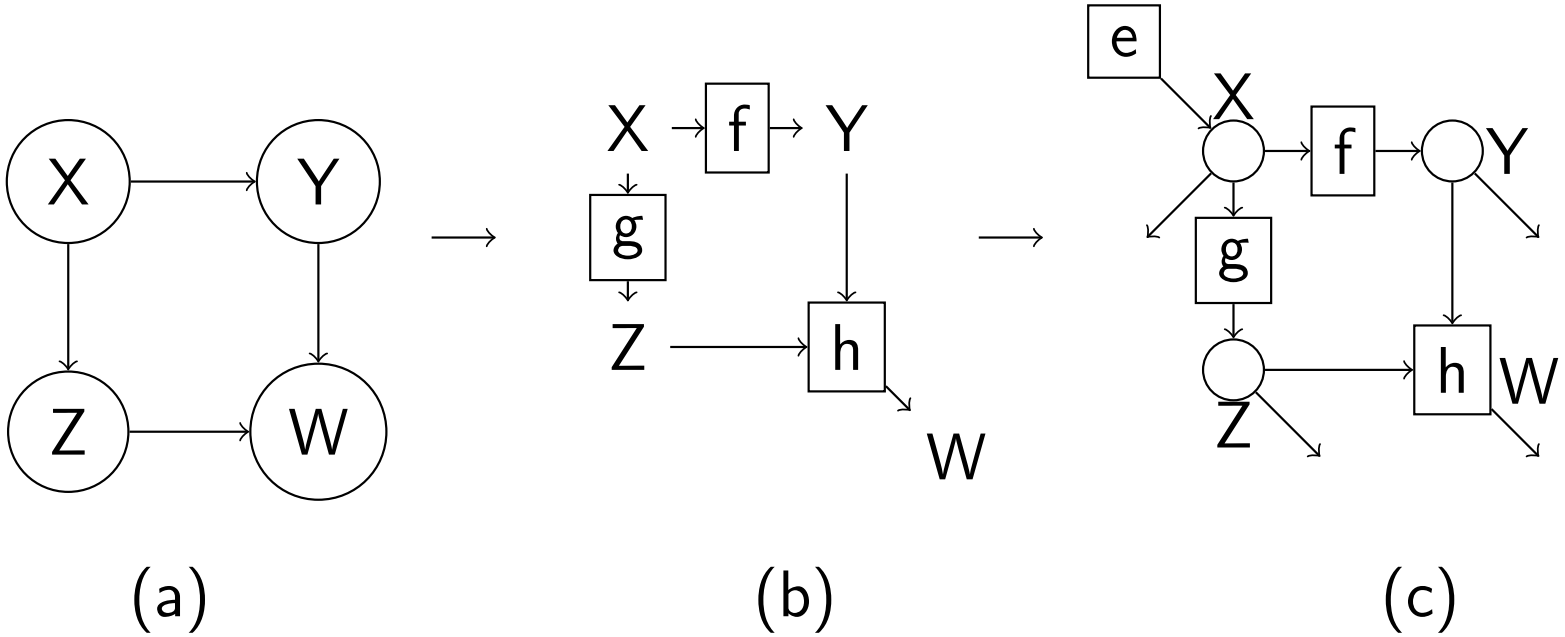
braid isomorphism
$b_{A\otimes B}$



symmetry relation

From graphical model to string diagram

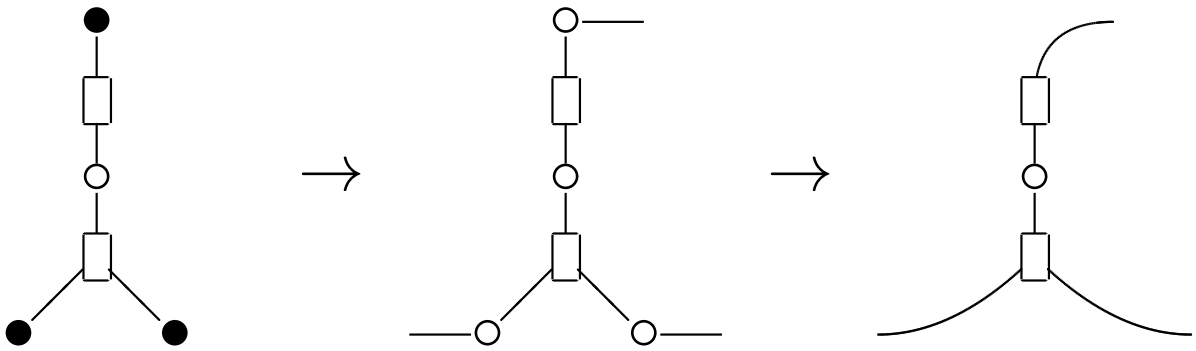
Q: What does the graphical language of certain monoidal categories with additional axioms look like?
A*: **Factor graph models**.



(a) (b) (c)

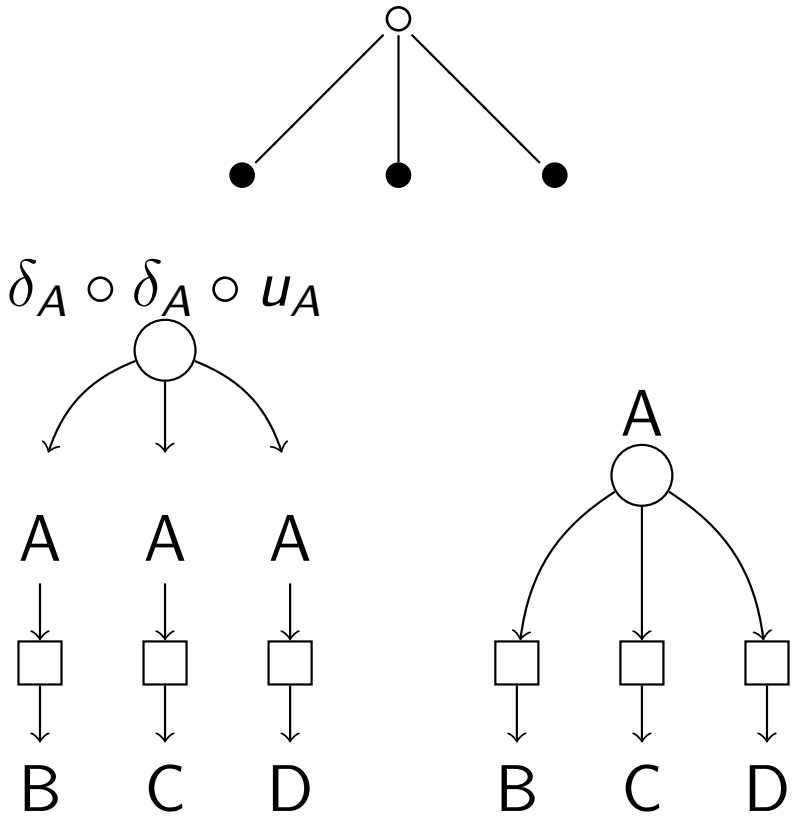
Converting a Bayesian network (a) to a directed factor graph (b) and a string diagram (c). Factor f is the conditional distribution $p_{y|x}$, g is $p_{z|x}$, and h is $p_{w|z,y}$.

From graphical model to string diagram

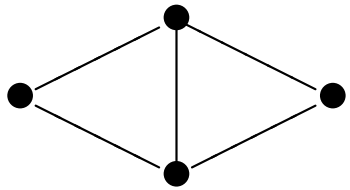


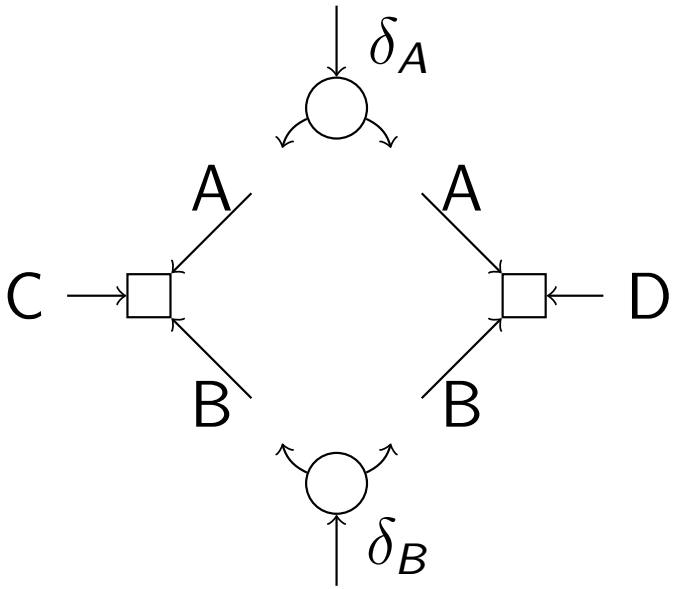
Converting an undirected factor graph to a string diagram.

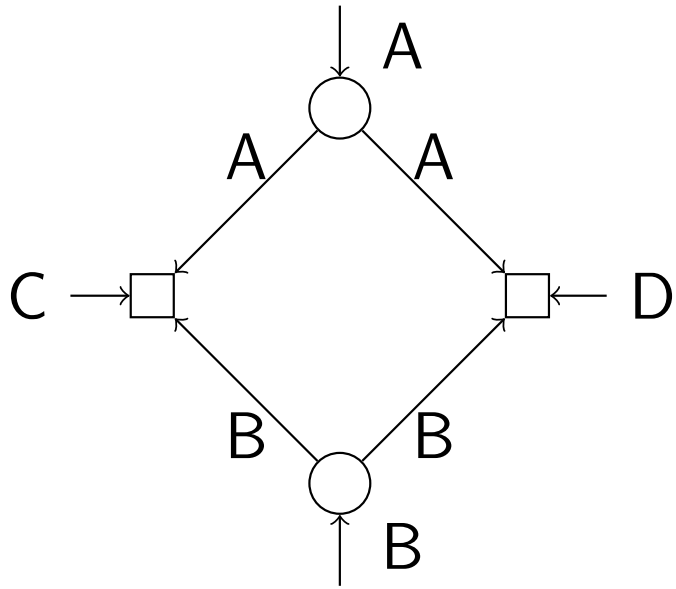
Example: naïve Bayes



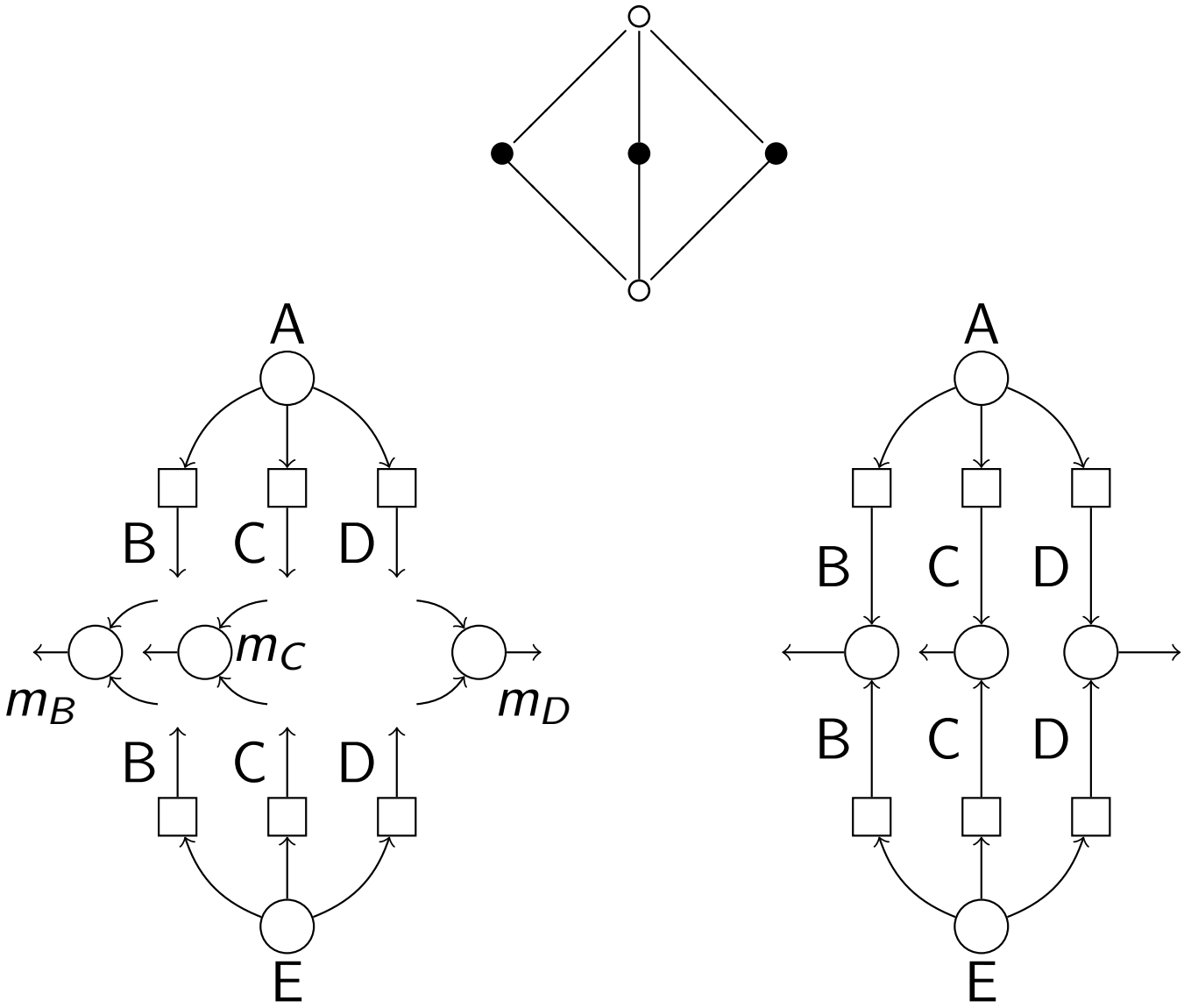
Example: gluing two triangles







RBM: Hadamard product of naïve Bayes



Graphical models as tensor networks

Roughly

- Wires are variables (Frobenius algebras)
- Boxes are factors at fixed parameters, or spaces of factors as parameters vary
- Under suitable assumptions
  - global properties
  - (such as the set of equations cutting out the space of representable probability distributions)
  - can be computed by gluing local properties

How do we describe these spaces of representable probability distributions?

Algebraic geometry of ~~tensor networks~~

1. Algebraic geometry of tensor networks
   - Tensors
   - Tensor Networks
   - Algebraic geometry
2. Algebraic Description of Graphical Models
   - Review of GM Defininitions
   - Algebraic and semialgebraic descriptions
   - Restricted Boltzmann machines
3. Identifiability, singular learning theory, other perspectives
   - Identifiability
   - Singular Learning Theory

What is algebraic geometry?

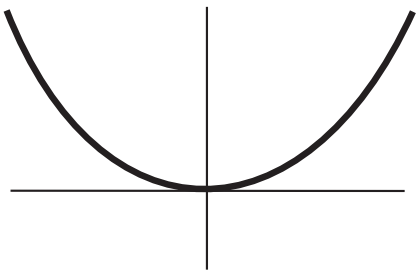
Study of solutions to systems of polynomial equations

- Consider the ring of multivariate polynomials $f \in \mathbb{C}[x_1, \ldots, x_n]$, e.g. $3x_2^2x_4 - 5x_3^3$
- Any polynomial $f \in \mathbb{C}[x_1, \ldots, x_n]$ has a zero locus

$$\{v = (v_1, \ldots, v_n) \in \mathbb{C}^n : f(v) = 0\}.$$

- This is the variety $V(f)$ cut out by f. For one polynomial, this variety is a hypersurface.

$V(\{p_2 - p_1^2\}) =$



What is algebraic geometry?

Study of solutions to systems of polynomial equations

- Consider the ring of multivariate polynomials $f \in \mathbb{C}[x_1, \ldots, x_n]$, e.g. $3x_2^2x_4 - 5x_3^3$
- Any polynomial $f \in \mathbb{C}[x_1, \ldots, x_n]$ has a zero locus

$$\{v = (v_1, \ldots, v_n) \in \mathbb{C}^n : f(v) = 0\}.$$

- This is the variety $V(f)$ cut out by f. For one polynomial, this variety is a hypersurface.
- For example, the set of probability distributions that can be represented by an RBM with 4 visible and 2 hidden nodes is part of a hypersurface.
  - its f has degree 110 and probably around 5 trillion monomials [Cueto-Yu]

Varieties defined by sets of polynomials

- Now suppose you have a set $\mathcal{F}$ of polynomials: a system of polynomial equations
- Requiring them all to hold simultaneously means we keep only the points where they all vanish:
  - the intersection of their hypersurfaces
- The zero locus

$$\{v = (v_1, \ldots, v_n) \in \mathbb{C}^n : f(v) = 0 \text{ for all } f \in \mathcal{F}\}$$

of a set of polynomials $\mathcal{F}$ is the variety $V(\mathcal{F})$.

Sets of polynomials defined by varieties

On the other hand,

- Given a set $S \subset \mathbb{C}^n$, the vanishing ideal of S is

$$I(S) = \{f \in \mathbb{C}[x_1, \ldots, x_n] : f(a) = 0 \ \forall a \in S\}.$$

- Hilbert's basis theorem: such an ideal has a finite generating set.

- Combining these operations:
  a set $S \subset \mathbb{C}^n$ has a Zariski closure $V(I(S))$.

Why applied algebraic geometry

- "Any sufficiently advanced field of mathematics can model any problem."
  - Algebraic geometry is as old as it gets.
- More formally:
  - Ideal membership / computing Gröbner bases is EXPSPACE-complete.
  - Matiyasevich's theorem: every recursively enumerable set is a Diophantine set (10th)
- So any question with a computable answer can be phrased in terms of algebraic geometry.
  - If done well, get geometric insight into the problem and deep hammers.

Implicitization

- We would like to study e.g. the space of probability distributions representable by various deep learning models.
- These models are not given as sets of polynomials $\mathcal{F}$
- They are given in terms of a parameterization described by a graph: each box in the tensor network is allowed to be a certain restricted set of tensors
- So we need to study the map from parameter space to probability space, its fibers, and its image.
- A complete description tells us what equations and what inequalities a probability distribution must satisfy in order to come from, say, a DBN

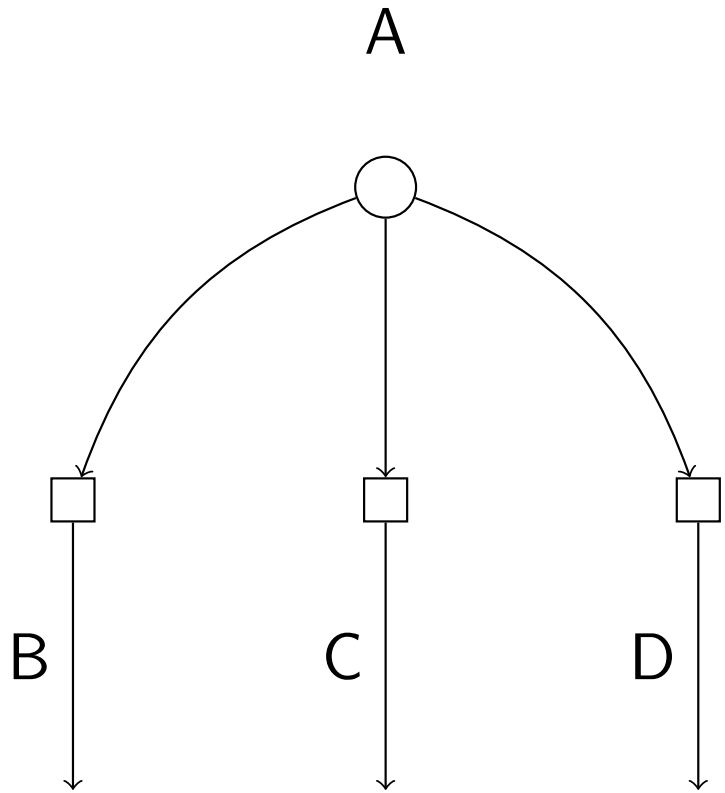
Implicitization

- Define a polynomial map ϕ from a parameter space $\Theta \subset \mathbb{C}^n$ to an ambient space $\mathbb{C}^m$

$$\begin{aligned} x &= t \\ y &= t^2 \end{aligned}$$

- Defines an image $\phi(\Theta) \subset \mathbb{C}^m$. What equations define, or cut out this set? $y - x^2 = 0$ cuts out the image.
- We took a Zariski closure
- The process of finding defining equations of the image is called implicitization
- This is hard! But not impossible.

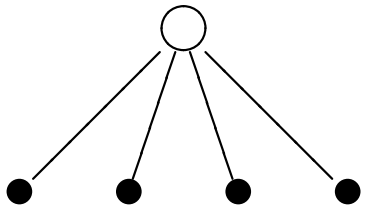
Example: Mixture of products



/ Naïve Bayes / Secant Segre / Tensor Rank

$\mathbb{P}^1$

$\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \hookrightarrow \mathbb{P}^{15}$
Segre variety defined by
2×2 minors of flattenings
of $2 \times 2 \times 2 \times 2$ tensor



$\sigma_2(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1)$
First secant of Segre variety
3×3 minors of flattenings

Dimension, equations defining such models?

Computational Algebraic Geometry

- There are computational tools for algebraic geometry, and many advances mix computational experiments and theory.
- Gröbner basis methods power general purpose software: Singular, Macaulay 2, CoCoA, (Mathematica, Maple)
  - Symbolic term rewriting
- Polyhedral methods (e.g. polymake, gfan) for certain problems; e.g. using work by Cueto and Yu, Bray and M- reduce implicitization to (very) large-scale linear algebra: just find the (1d) kernel of a 5 trillion $\times$ 5 trillion matrix for $RBM_{4,2}$.
- Computational algebraic geometry computations now routinely burn millions of CPU-hours of cluster compute time (e.g. implicitization, searching for efficiently contractable networks).

Numerical Algebraic Geometry

- Salmon Problem: Determine the ideal defining the fourth secant variety of $\mathbb{P}^3 x \mathbb{P}^3 x \mathbb{P}^3$. Set theoretic [Friedland 2010], further progress [Bates, Oeding 2010], [Friedland, Gross 2011].
- Numerical Algebraic Geometry: Numerical methods for approximating complex solutions of polynomial systems.
  - Homotopy continuation (numerical path following).
  - Can be used to find isolated solutions or points on each positive-dimensional irreducible component.
  - Can scale to thousands of variables for certain problems.
  - Reliable, parallelized, adaptive multiprecision software is available: *Bertini* (Bates, Hauenstein, Sommese, and Wampler).

Why are geometers interested?

- Applications (especially tensor networks in statistics and CS) have revived classical viewpoints such as invariant theory.
- Re-climbing the hierarchy of languages and tools (Italian school, Zariski-Serre, Grothendieck) as applied problems are unified and recast in more sophisticated language.
- Applied problems have also revealed gaps in our knowledge of algebraic geometry and driven new theoretical developments and computational tools
  - Objects which are "large": high-dimensional, many points, but with many symmetries
  - These often stabilize in some sense for large n.

Lectures 2 and 3

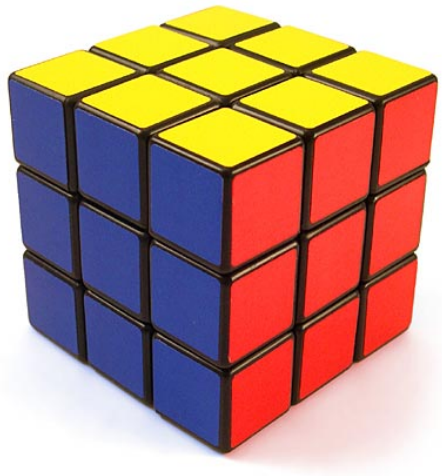
Last time

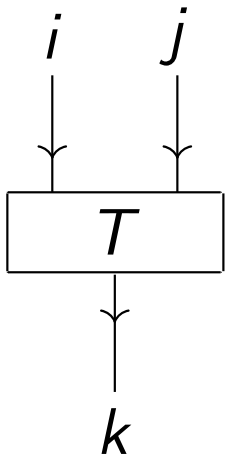
Last time, we talked about the

algebraic geometry of tensor networks

and how to learn something about this geometry

Tensors





A tensor $T = (t_{ijk})$

- represents a multilinear transformation $U \otimes V \to W$, $W \to U \otimes V$, $U \otimes W \to V$, etc.
- is a multi-way array (here 3-way)
- with a multilinear action on each "leg" or "side"

Tensor decomposition is possible but more subtle

Data arrives in tensor format, and something is lost by flattening

One way things get harder with tensors

- Which tensor products $\mathbb{C}^{d_1} \otimes \cdots \otimes \mathbb{C}^{d_n}$ have finitely many orbits under $\mathrm{GL}(d_1, \mathbb{C}) \times \cdots \times \mathrm{GL}(d_n, \mathbb{C})$?
- The answer for matrices is easy
- Kac (1980), Parfenov (1998, 2001): up to $\mathbb{C}^2 \otimes \mathbb{C}^3 \otimes \mathbb{C}^6$, orbit representatives and abutment graph

Case $(2, m, n)$	The number of orbits of $\mathrm{GL}_2 \times \mathrm{GL}_m \times \mathrm{GL}_n$	$\deg f$
$(2,2,2)$	7	4
$(2,2,3)$	9	6
$(2,2,4)$	10	4
$(2,2,n),\ \ n \geqslant 5$	10	0
$(2,3,3)$	18	12
$(2,3,4)$	24	12
$(2,3,5)$	26	0
$(2,3,6)$	27	6
$(2,3,n),\ \ n \geqslant 7$	27	0

Another way

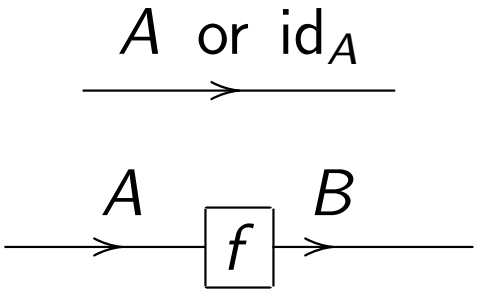
- Consider a matrix turned into a vector

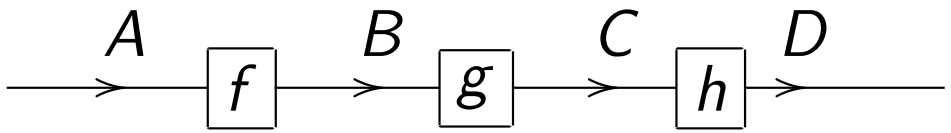
$$[x_1, x_2, x_3, \ldots x_\ell]$$

- can you compute its rank, SVD, kernel, etc?

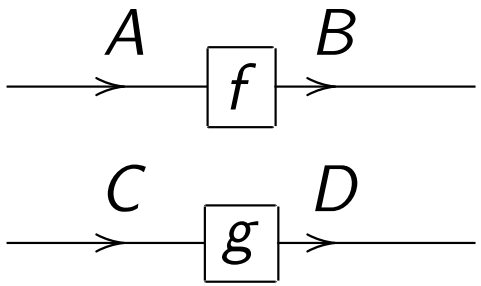
Composition

- Connect wires to compose/contract: $h \circ g \circ f$



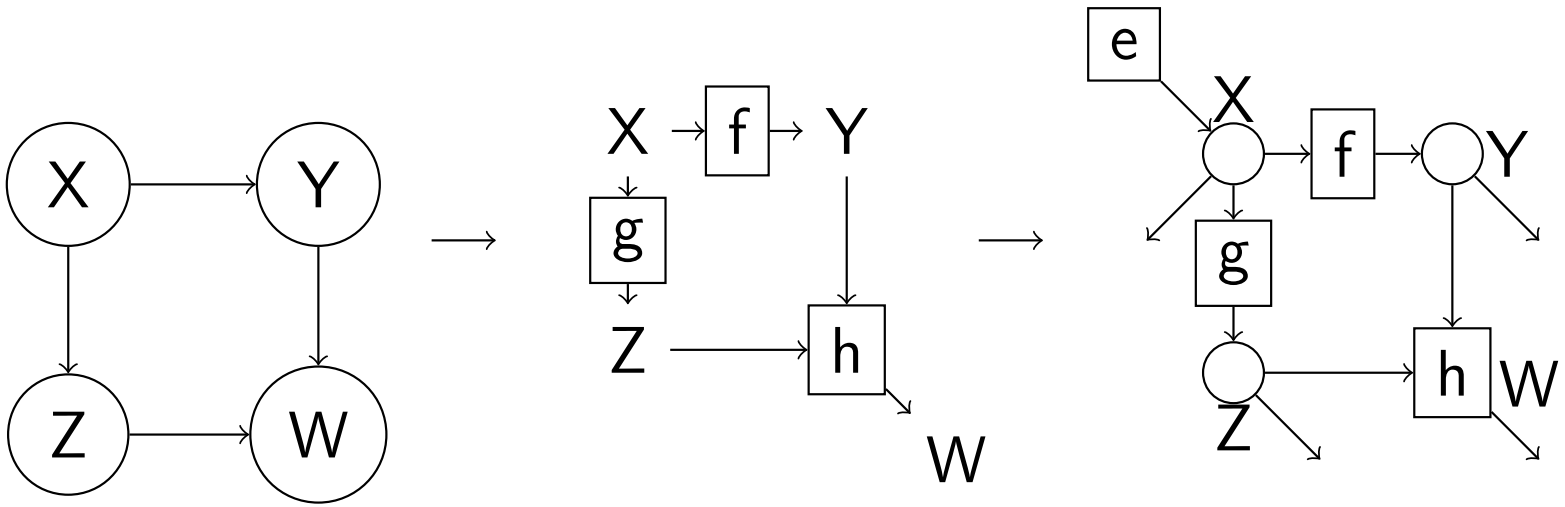


- juxtpose to tensor/run in parallel $f \otimes g$



From these primitives (and duals, swaps, special maps) can build complex networks such as graphical models...

Tensor networks



Algebraic geometry

Study of solutions to systems of polynomial equations

- Ring of multivariate polynomials $f \in \mathbb{C}[x_1, \ldots, x_n]$, e.g. $3x_2^2x_4 - 5ix_3^3$
- The zero locus

$$\{v = (v_1, \ldots, v_n) \in \mathbb{C}^n : f(v) = 0 \text{ for all } f \in \mathcal{F}\}$$

  of a set of polynomials $\mathcal{F}$ is the variety $V(\mathcal{F})$.
- Given a set $S \subset \mathbb{C}^n$, the vanishing ideal of S is

$$I(S) = \{f \in \mathbb{C}[x_1, \ldots, x_n] : f(a) = 0 \; \forall a \in S\}.$$

  Hilbert's basis theorem: such an ideal has a finite generating set.
- A set $S \subset \mathbb{C}^n$ has a Zariski closure $V(I(S))$.

Implicitization

- Define a polynomial map ϕ from a parameter space $\Theta \subset \mathbb{C}^n$ to an ambient space $\mathbb{C}^m$

$$\begin{aligned} x &= t \\ y &= t^2 \end{aligned}$$

- Defines an image $\phi(\Theta) \subset \mathbb{C}^m$. What equations define, or cut out this set? $y - x^2 = 0$ cuts out the image.
- We took a Zariski closure
- The process of finding defining equations of the image is called implicitization

Algebraic geometry and tensor networks

To an algebraic geometer, a tensor network

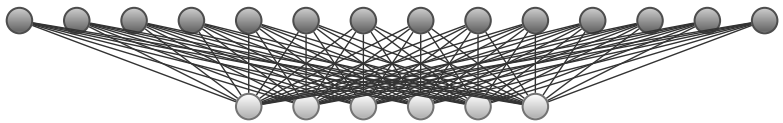
- appearing in machine learning (statistics, signal processing, computational complexity, quantum computation, ...)
- describes a regular map ϕ from the parameter space (choice of tensors at the nodes) to an ambient space.
- The image of ϕ is an algebraic variety of representable probability distributions,
- The fibers tell us about identifiability, transferability, and learning rate

1. Algebraic geometry of tensor networks
   - Tensors
   - Tensor Networks
   - Algebraic geometry

2. Algebraic Description of Graphical Models
   - Review of GM Defininitions
   - Algebraic and semialgebraic descriptions
   - Restricted Boltzmann machines

3. Identifiability, singular learning theory, other perspectives
   - Identifiability
   - Singular Learning Theory

Algebraic description of probabilistic models

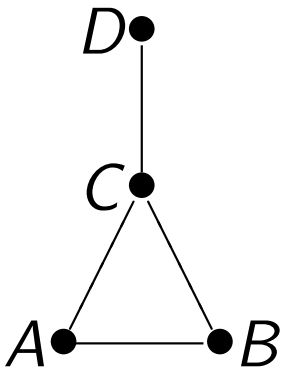


"Statistical models are algebraic varieties"

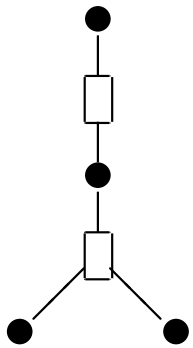
- What distributions can be represented by a (graphical) model?
- What is the geometry of the parameterization map?
- Implications for approximation and optimization (learning) performance?

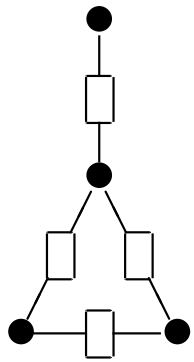
Hierarchical models, undirected graphical models

- Model joint probability distributions on N random variables $X_1, \ldots, X_N$ with finitely many states $d_1, \ldots, d_N$.
- Define dependence locally by a simplicial complex on $\{1, 2, \ldots, N\}$, parameterizing a family of probability distributions by potential functions or factors, one per maximal simplex.
- In an undirected graphical model, the simplicial complex is the clique complex of an undirected graph.



UGM


FG


FG

Hierarchical models: probability distribution

- Model joint probability distributions on N random variables $X_1, \ldots, X_N$ with finitely many states $d_1, \ldots, d_N$.
- Define dependence locally by a simplicial complex on $\{1, 2, \ldots, N\}$, parameterizing a family of probability distributions by potential functions or factors, one per maximal simplex.
- In an undirected graphical model, the simplicial complex is the clique complex of an undirected graph.

Defines a family of probability distributions on discrete random variables $X_1, \ldots, X_N$, where X_i has d_i states by (before marginalizing)

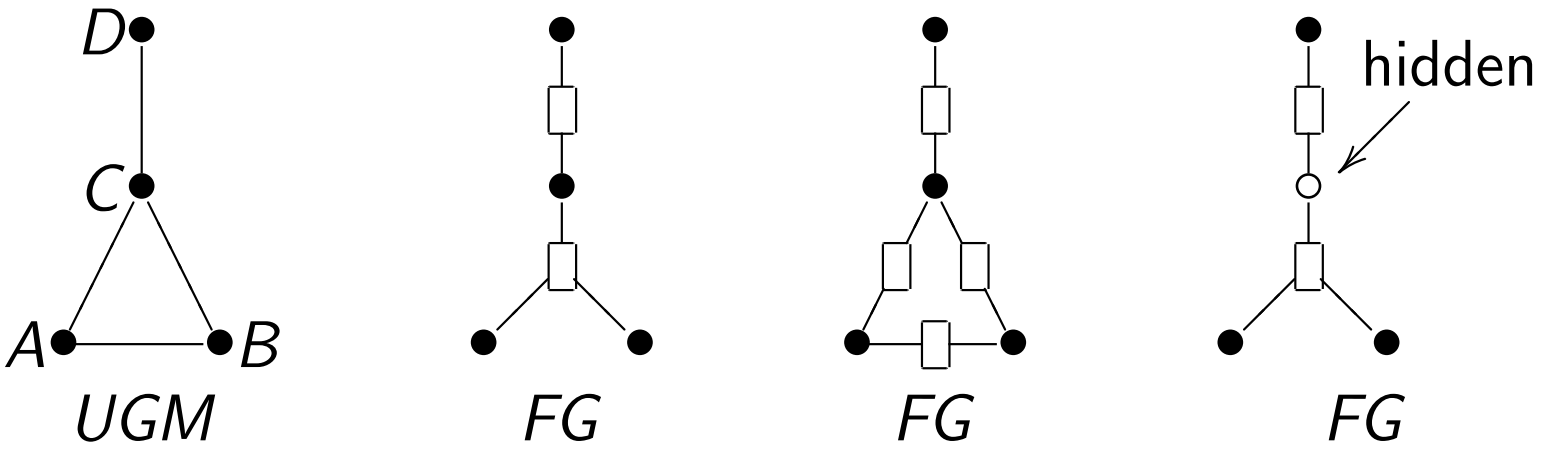
$$p_{\mathbf{M}}(x) = \frac{1}{Z} \prod_{s \in S} \Xi_s(x_s)$$

where x_s is the state vector restricted to the vertices in s, each Ξ_s is a tensor corresponding to the factor associated to simplex s.

Hierarchical models: undirected factor graph models

A hierarchical model defines a factor graph, which is a bipartite graph $\Gamma = (V \cup H, F, E)$ of nodes, factors and edges.

- Each $i \in \{1, \ldots, N\} = H \cup V$ is labeled by a random variable X_i and denoted by a open $\circ$ or filled $\bullet$ disc according to whether it is Hidden (latent) or Visible respectively.
- Each maximal simplex $s \in S$ is denoted by a box $\boxed{f_s}$ labeled by a factor $f_s \in F$, and is connected by an edge to each variable disc X_i where $i \in s$.



Implicitization and graphical models

- The Hammersley-Clifford Theorem is a theorem about implicitizing undirected graphical models
- They delayed publication for years trying to address the nonnegative case
- This was completed in [Geiger, Meek, Sturmfels 2006] by studying the algebraic geometry of these models ("Algebraic Statistics")

Bayesian networks: directed factor graph models

A (discrete) Bayesian network $\mathbf{M} = (H, V, d, G)$ is based on a directed acyclic graph G.

- Vertices partitioned $[N] = H \cup V$ into hidden and visible variables; each variable $i \in [N]$ has a number d_i of states.
- The parameterization defines for each variable x a conditional probability distribution (a singly stochastic matrix) $p(x_i|x_{\mathrm{pa}(i)})$ where $\mathrm{pa}(i)$ is the set of vertices which are parents of i.
- Then

$$p_{\mathbf{M}}(v) = \sum_{x_H} \prod_{i \in [N]} p(x_i|x_{\mathrm{pa}(i)})$$

with $p(x_i|x_{\mathrm{pa}(i)}) \geq 0$, $\sum_i p(x_i|x_{\mathrm{pa}(i)}) = 1$ and no global normalization is needed because of the local normalization.

Bayesian networks: directed factor graph models

- Every Bayesian network can be written as a directed factor graph model, but not conversely [Frey 2003].
- Algebraic geometry of Bayesian networks [Garcia, Stillman, Sturmfels 2005]

"Unhidden" Binary Deep Belief Network

Consider a binary DBN with layer widths $n_0, \ldots, n_\ell$. An "unhidden" binary DBN defines joint probability distributions of the form

$$\begin{aligned} P(h^0, h^1, \ldots, h^\ell) &= P(h^{\ell-1}, h^\ell) \prod_{k=0}^{\ell-2} P(h^k | h^{k+1}) \,, \\ P(h^k | h^{k+1}) &= \prod_{j=1}^{n_k} P(h_j^k | h^{k+1}) \,, \\ P(h_j^k | h^{k+1}) &\propto \exp\left(h_j^k b_j^k + h_j^k \sum_{i=1}^{n_{k+1}} W_{j,i}^{k+1} h_i^{k+1} \right) \,, \end{aligned}$$

- $h^k = (h_j^k)_j \in \{0,1\}^{n_k}$ is the state of the units in the kth layer,
- $W_{j,i}^k \in \mathbb{R}$ is the *connection weight* between the units j and i from the $(k-1)$th and kth layer respectively, and
- $b_j^k \in \mathbb{R}$ is the *bias weight* of the jth unit in the kth layer.

Binary DBN

- Now the *DBN model* $\mathrm{DBN}(n_0, n_1, \ldots, n_\ell)$ is the set of marginal distributions

$$P(h^0) = \sum_{h^1 \in \{0,1\}^{n_1}, \ldots, h^\ell \in \{0,1\}^{n_\ell}} P(v, h^1, \ldots, h^\ell), \qquad h^0 \equiv v \in \{0,1\}^{n_0} \tag{1}$$

  of joint probability distributions of that form.
- The DBN has

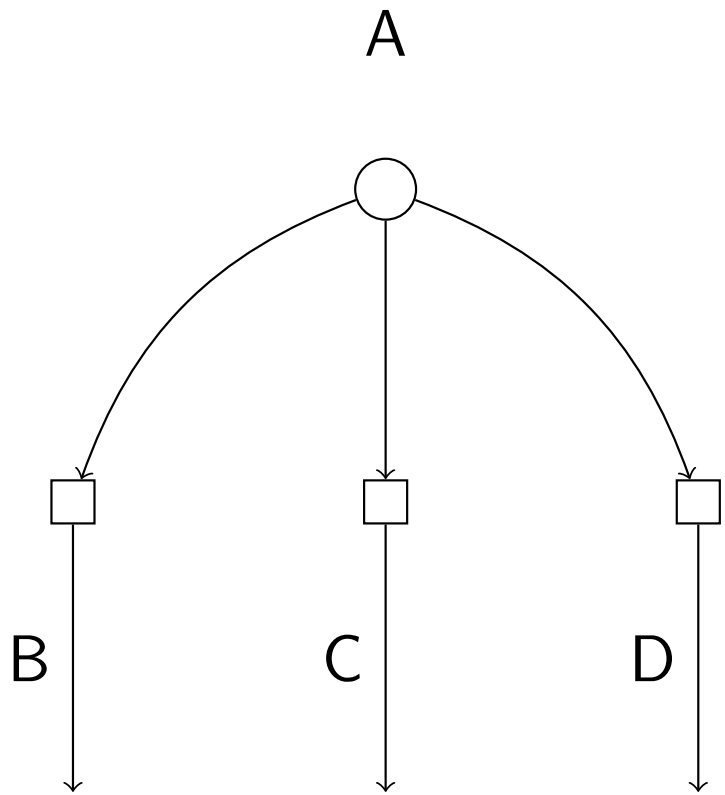
$$d = (\sum_{k=1}^{\ell} n_{k-1} n_k) + (\sum_{k=0}^{\ell} n_k) \text{ parameters.}$$

- So this is its expected dimension if there is no collapse or waste of parameters. Which tuples $(n_0, \ldots, n_\ell)$ have the expected dimension?

Where we are

- The DBN contains many of the known models, hence
- Don't have a complete semialgebraic description of DBN, DBM.
- Have partial information about representational power
- Have algebraic, semialgebraic descriptions of submodels: naïve Bayes, HMM, trees, RBM, etc.
  - In some cases (especially small number of states), this is done
  - In others, just have coarse information (dimension, relative power)
- Translating that understanding to something prescriptive is ongoing
- Let's look at some of these submodels

Naïve Bayes / Secant Segre / Tensor Rank



Naïve Bayes / Secant Segre / Tensor Rank

Look at one hiden node in such a network, binary variables

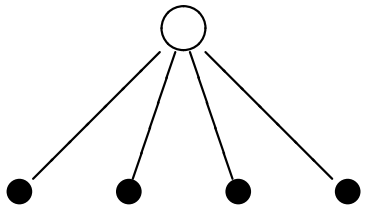
$\mathbb{P}^1$

$\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \hookrightarrow \mathbb{P}^{15}$

Segre variety defined by

2×2 minors of flattenings

of $2 \times 2 \times 2 \times 2$ tensor



$\sigma_2(\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1)$

First secant of Segre variety

3×3 minors of flattenings

Dimension, equations defining such models?

Expected dimension of secant varieties

- The expected dimension of $\sigma_k(\mathbb{P}^1)^n$ is

$$\min(kn + k - 1, 2^n - 1)$$

  (e.g. by a parameter count)
- But (especially for small n) things can collide and we can get a defect, where the dimension is less than expected
- Dimension is among the first questions one can ask about a variety
  - Is there hope for identifiability?
  - are there wasted parameters/ positive-dimensional fibers?
  - how big a model is needed to be able to represent all distributions?

Dimension of secant varieties

- Recently [Catalisano, Geramita, Gimigliano 2011] showed $\sigma_k(\mathbb{P}^1)^n$ has the expected dimension

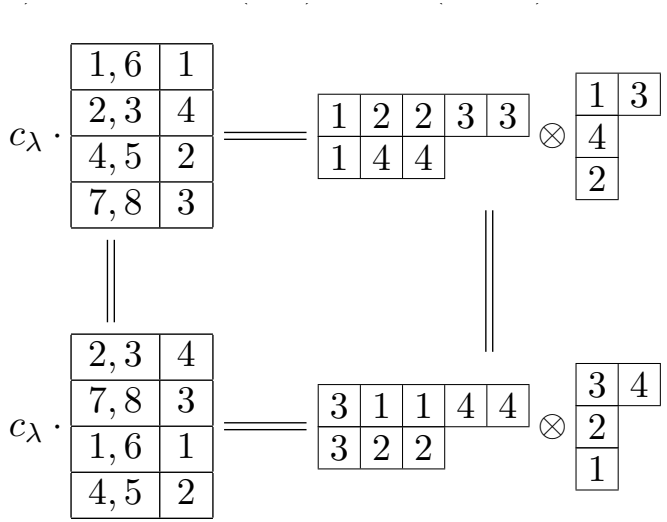
$$\min(kn + k - 1, 2^n - 1)$$

  except $\sigma_3(\mathbb{P}^1)^4$ where it is 13 not 14.
- Progress in Palatini 1909, . . . , Alexander Hirschowitz 1995, 2000, CGG 2002,03,05, Abo Ottaviani Peterson 2006, Draisma 2008, others.
- Classically studied, revived by applications to statistics, quantum information, and complexity; shift to higher secants, solution.
- So a generic tensor of $(\mathbb{C}^2)^{\otimes n}$ can be written as a sum of $\lceil \frac{2^n}{n+1} \rceil$ decomposable tensors, no fewer.

Representation theory of secant varieties

Raicu (2011) proved the ideal-theoretic GSS [Garcia Stillman Sturmfels 05] conjecture

- Equations defining $\sigma_2(\mathbb{P}^{k_1} \times \cdots \times \mathbb{P}^{k_n})$
- Using representation theory of ideal of $\sigma_2(\mathbb{P}^{k_1} \times \cdots \times \mathbb{P}^{k_n})$ as a $GL_{k_1} \times \cdots \mathrm{GL}_{k_n}$-module
- (progress in [Landsberg Manivel 04, Landsberg Weyman 07, Allman Rhodes 08]).



Equations of the naïve Bayes model (secant varieties of Segre varieties)

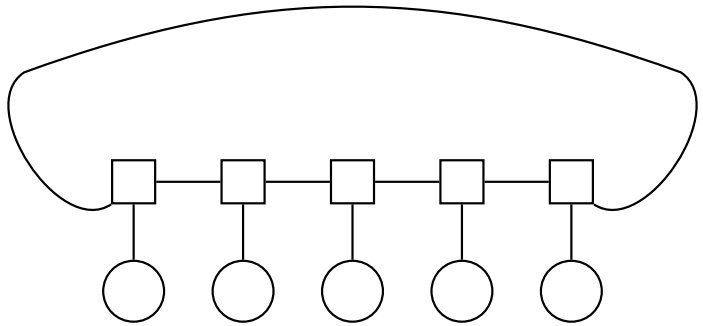
Good news/bad news

- Good News: We know them for small number of states
- Bad News: But we don't know them for large numbers of states.
- Good News: In some cases, just minors of flattenings
- Bad News: But not in general.
- Good News: Many models (trees, RBM, DBN) are built by gluing naïve Bayes models together, so we have some information and the good news above propagates
- Bad News: But so does the bad news.

Algebraic description of Hidden Markov Models

A simplified (circular) version. Fix parameter matrices $A_1, \ldots, A_d$. Then up to a global rescaling,

$$p = \sum_{i_1,\ldots,i_n} \mathrm{tr}(A_{i_1} \cdots A_{i_n}) e_{i_1 i_2 \cdots i_n}$$



Algebraic description of Hidden Markov Models

A simplified (circular) version. Fix parameter matrices $A_1, \ldots, A_d$. Then up to a global rescaling,

$$p = \sum_{i_1,\ldots,i_n} \mathrm{tr}(A_{i_1} \cdots A_{i_n}) e_{i_1 i_2 \cdots i_n}$$

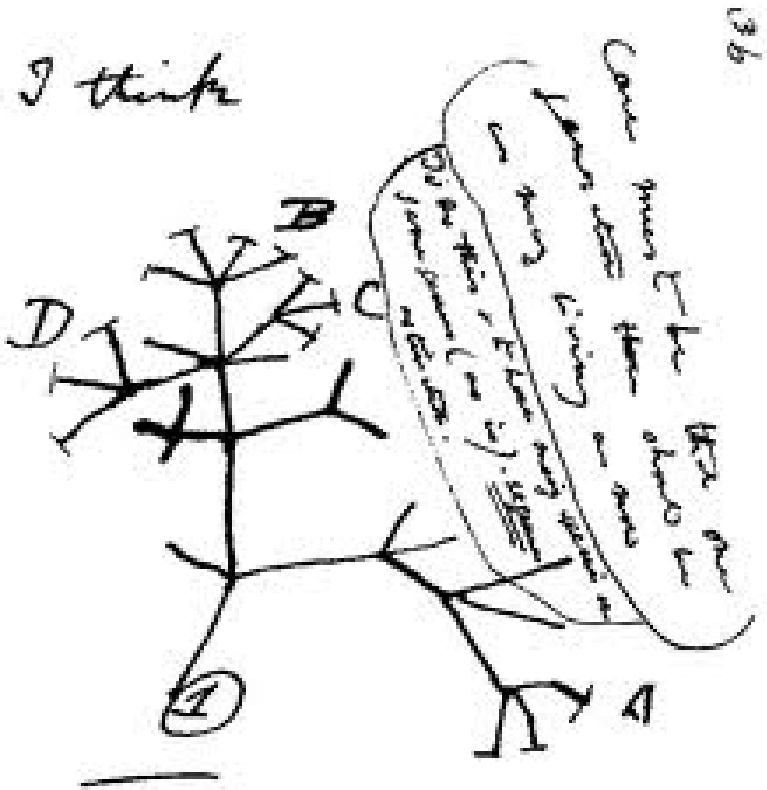
What are the polynomial relations that hold among the coefficients

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

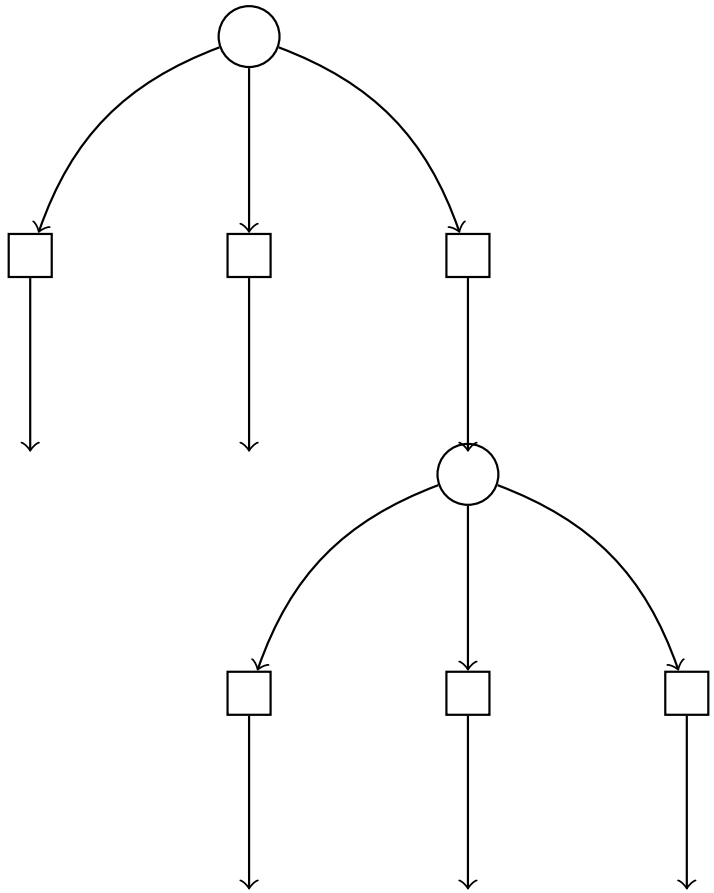
That is, the ideal $I = \{f : f(p_{i_1,\ldots i_n}) = 0\}$ of polynomials f in the coefficients such that $f(p_{i_1,\ldots i_n}) = 0$.

Series of papers [Bray and M- 2006], [Schönhuth 2008, 2011], [Critch 2012] provide characterizations, membership tests, identifiability, etc.

(Phylogenetic) Trees



Trees

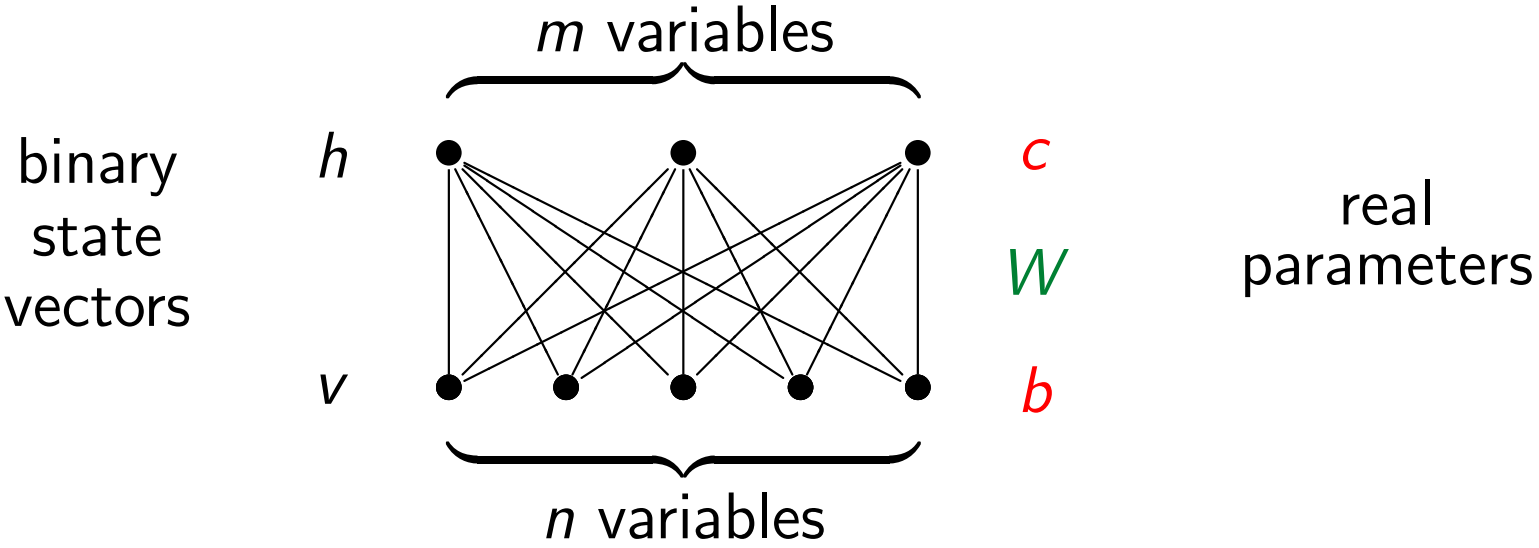


Trees and the General Markov Model

- Studied in a long series of papers by authors including Sturmfels-Sullivant, Casanellas, Draisma-Kuttler, Allman-Rhodes, many others
  - Many techniques were developed first on this reasonably tractable class
- Ideas include changes of coordinates (Fourier transform, cumulant coordinates), gluing constructions, hard work
- Now have complete algebraic description of many special classes
- Complete semi-algebraic description of GMM for small number of states [Allman et al. 2012]
- This is a submodel of the DBN.

Restricted Boltzmann Machines

pre-RBM: graphical model on a bipartite graph



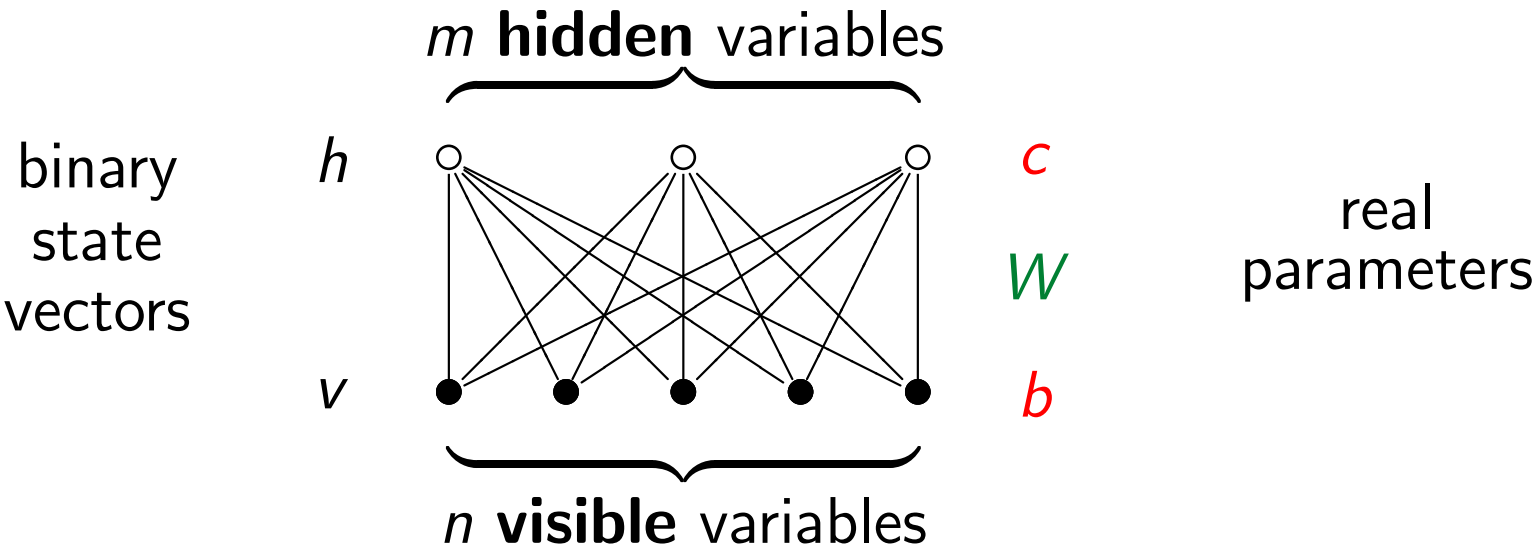
Unnormalized potential is built from node and edge parameters

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

The probability distribution on the binary random variables is

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

Restricted Boltzmann machines



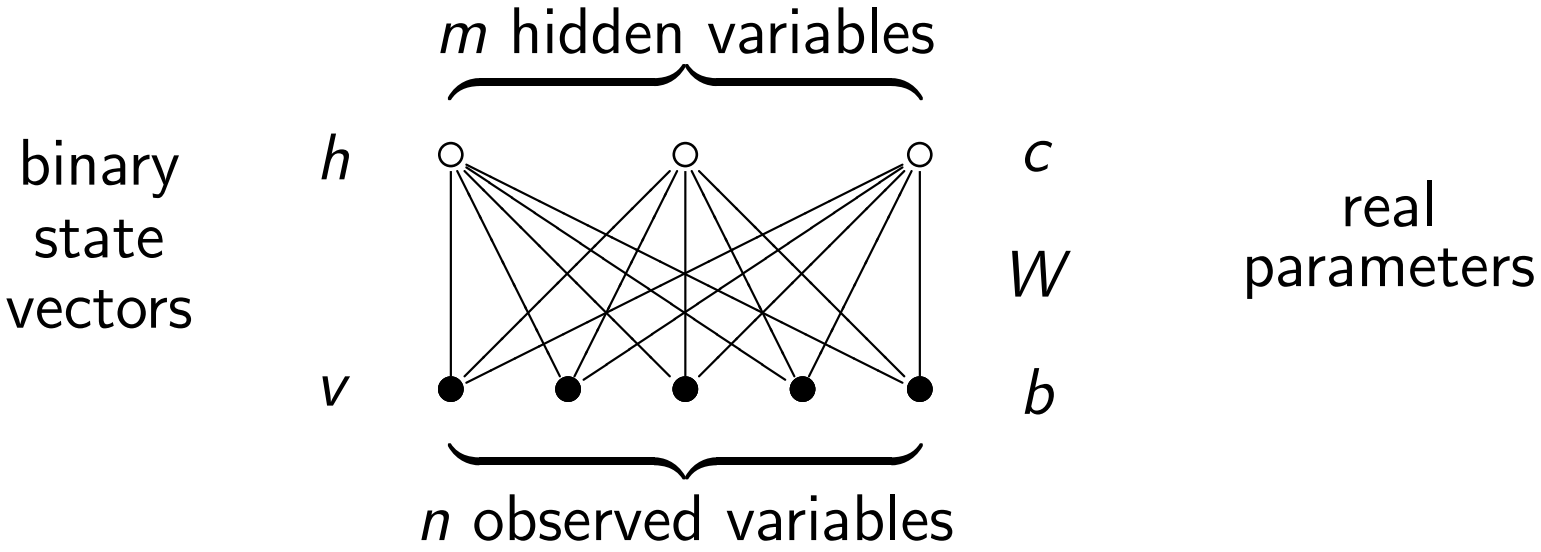
Unnormalized fully-observed potential is

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

The probability distribution on the visible random variables is

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

Restricted Boltzmann machines



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

Hadamard Products of Varieties

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

$$X \times Y \dashrightarrow \mathbb{P}^\ell, \ (x, y) \mapsto (x_0 y_0 : x_1 y_1 : \ldots : x_\ell y_\ell).$$

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

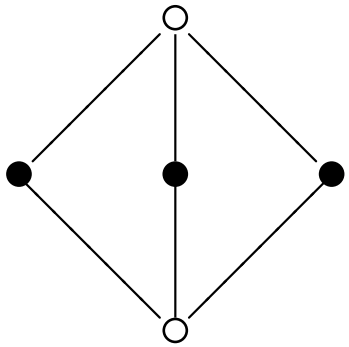
If M is a subset of the simplex $\Delta_{\ell-1}$ then $M^{[m]}$ is also defined by componentwise multiplication followed by rescaling so that the coordinates sum to one. This is compatible with taking Zariski closure: $\overline{M^{[m]}} = \overline{M}^{[m]}$

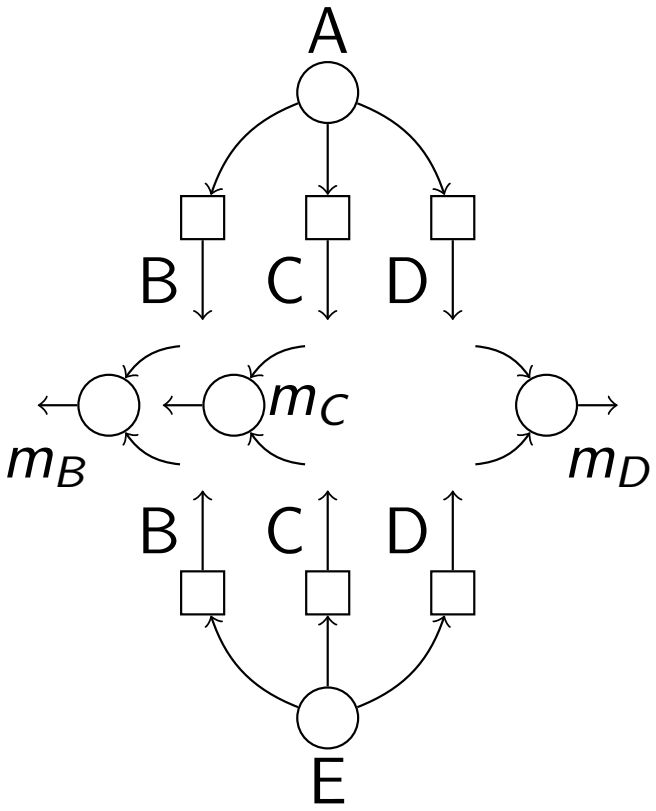
Lemma

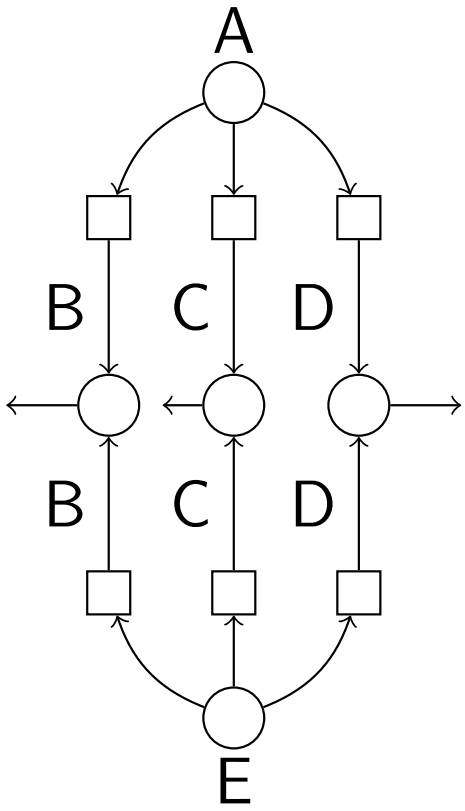
RBM variety and RBM model factor as

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

RBM as Hadamard product of naïve Bayes







Representational power of RBMs

Conjecture

The restricted Boltzmann machine has the expected dimension.

That is, M_n^m *is a semialgebraic set of dimension* $\min\{nm + n + m, 2^n - 1\}$ *in* Δ_{2^n-1}.

Expected dimension

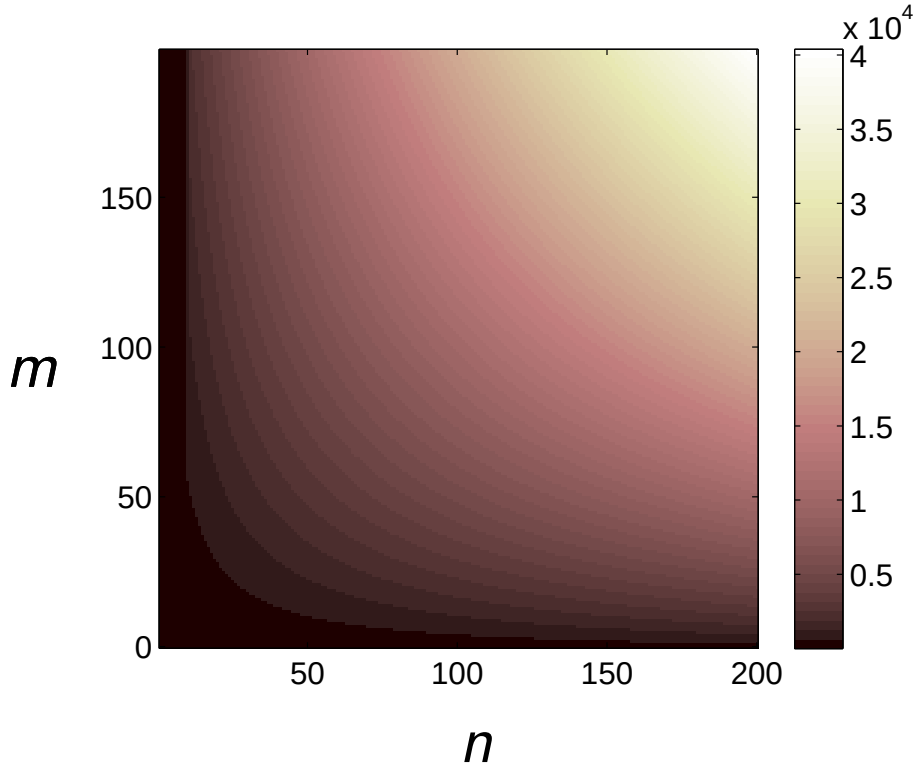


Figure: The expected dimension of the model $\mathrm{RBM}_{n,m}$, and of $\mathcal{M}_{n,m+1}$, $\min\{2^n - 1, nm + n + m\}$. There is a barely noticeable irregularity on the left side of the image, where the dimension of $\mathrm{RBM}_{n,m}$ equals the dimension of the ambient probability simplex $\mathcal{P}_n$ for all large enough m.

Representational power of RBMs

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

Theorem (Cueto M- Sturmfels)

The restricted Boltzmann machine has the expected dimension

- $nm + n + m$ *when* $m < 2^{n-\lceil \log_2(n+1) \rceil}$
- $\min\{nm + n + m, 2^n - 1\}$ *when* $m = 2^{n-\lceil \log_2(n+1) \rceil}$ *and*
- $2^n - 1$ *when* $m \geq 2^{n-\lfloor \log_2(n+1) \rfloor}$.

- Covers most cases of restricted Boltzmann machines in practice, as those generally satisfy $m \leq 2^{n-\lceil \log_2(n+1) \rceil}$.
- Proof uses tropical geometry, coding theory

Tropical RBM Model

- Tropical geometry is the "polyhedral shadow" of algebraic geometry.
- The process of passing from ordinary arithmetic to the max-plus algebra is known as *tropicalization*.
- The tropicalization Φ of our RBM parameterization is the map $\Phi : \mathbb{R}^{nm+n+m} \to \mathbb{TP}^{2^n-1} = \mathbb{R}^{2^n}/\mathbb{R}(1,1,\ldots,1)$ whose 2^n coordinates are the tropical polynomials

$$q_v \;=\; \max\left\{h^\top W v + b^\top v + c^\top h \;:\; h \in \{0,1\}^m\right\}$$

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

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

Tropical RBM Variety

- 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 RBM variety TV_n^m 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^m (or on M_n^m).

Understand tropical variety, use:

$$\dim(TM_n^m) \leq \dim(TV_n^m) = \dim(V_n^m) = \dim(M_n^m) \leq \min\{nm+n+m, 2^n-1\}$$

and coding theory to obtain the result.

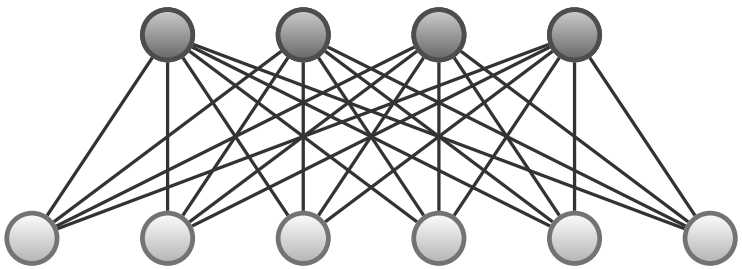
Relative representational power

- Another way to study the representational power of RBMs and DBNs is to compare them with other models
- When does one potential model contain another?

When does a mixture of products contain a product of mixtures?

Product of Mixtures (RBM)

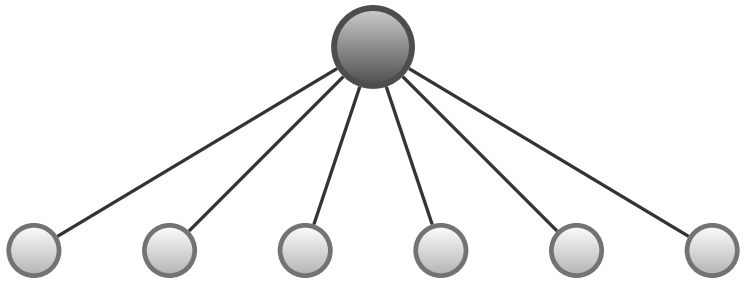
$\{0,1\}\{0,1\}\{0,1\}\{0,1\}$



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

Mixture of Products

$\{0,1,\ldots,k-1\}$



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

Problem

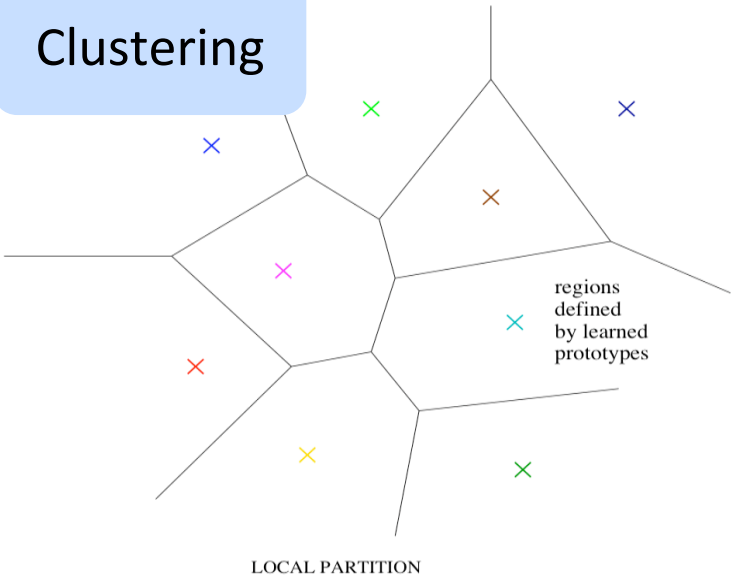
Given some $n, m \in \mathbb{N}$, what is the smallest $k \in \mathbb{N}$ for which the k-mixture of product distributions on n binary variables contains the RBM model with n visible and m hidden units?

Exponentially more efficient

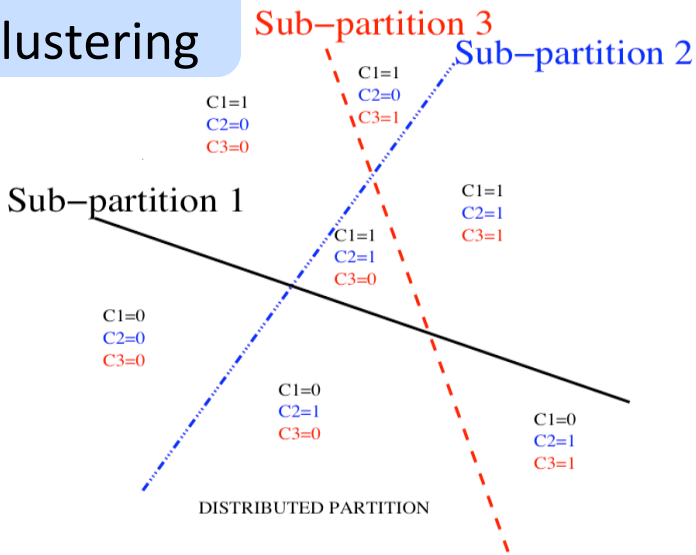
From Yoshua's first talk

#2 The need for distributed representations

Clustering



Multi-Clustering



Learning a **set of features** that are not mutually exclusive can be exponentially more statistically efficient than nearest-neighbor-like or clustering-like models

When does a mixture of products contain a product of mixtures?

- The number of parameters of the smallest mixture of products containing the RBM
  - grows exponentially in the number of parameters of the RBM
  - for any fixed ratio of hidden vs. visible units $0 < m/n < \infty$
- Such results aid our understanding of
  - how models complement each other,
  - why techniques such as deep learning can be expected to succeed, and
  - when model selection can be based on theory.

Not very often

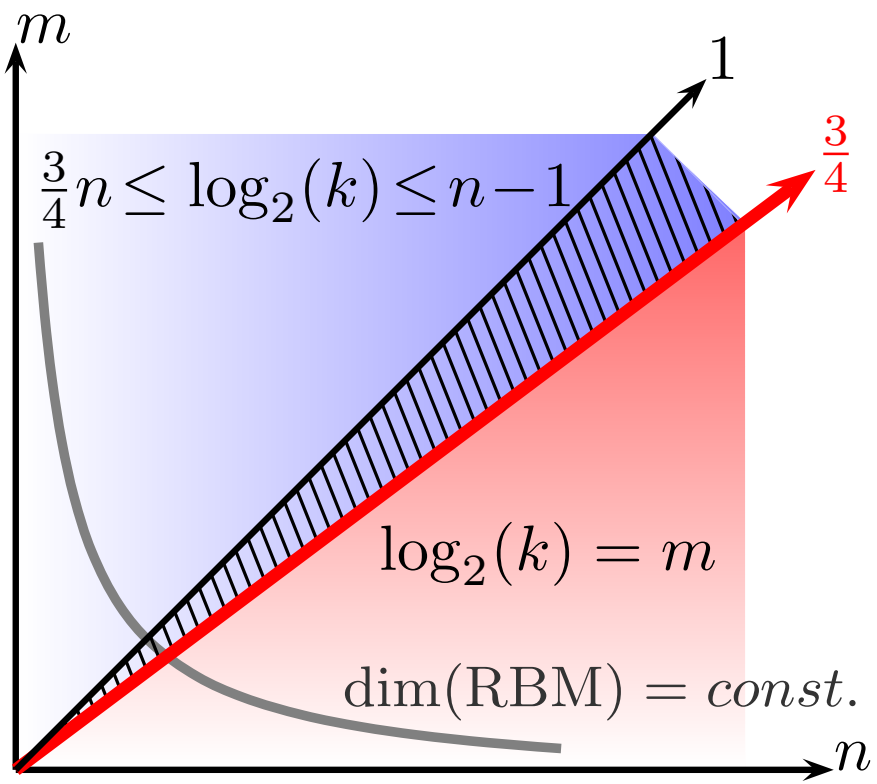


Figure: Plot of the smallest k for which $\mathcal{M}_{n,k}$ contains $\text{RBM}_{n,m}$. Fixing dimension (grey line), the RBMs which are hardest to represent as mixtures are those where $m/n \approx 1$.

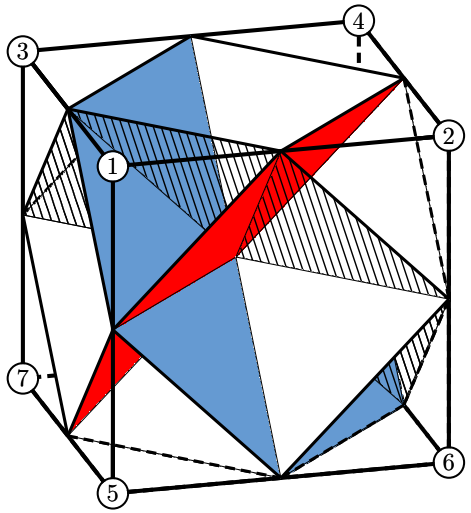
Modes and strong modes

Definition

Call $x \in \{0,1\}^n$ a *mode* of $p \in \mathcal{P}_n$ if $p(x) > p(\hat{x})$ for all $\hat{x}$ with $d_H(\hat{x}, x) = 1$, and a *strong mode* if $p(x) > \sum_{\hat{x}: d_H(\hat{x},x)=1} p(\hat{x})$.

- One way to make precise that the RBM can represent more complicated distributions than a mixture model of similar size is to study this bumpiness in Hamming space.
- On the one hand the sets of strong modes $C \subset \{0,1\}^n$ realizable by a mixture model $\mathcal{M}_{n,k}$ are exactly the binary codes of minimum Hamming distance two and cardinality at most k.
- On the other hand...

RBMs and Linear Threshold Codes



Definition

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

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

If the functions f_i can be chosen self-dual (hyperplanes are central), then $\mathcal{C}$ is called *homogeneous*.

Strong modes and linear threshold codes

- On the other hand,
- for codes $\mathcal{C} \subseteq \{0,1\}^n, |\mathcal{C}| = 2^m$ of minimum distance two,
- when $\mathcal{C}$ is a homogeneous linear threshold code (LTC)
- then $\text{RBM}_{n,m}$ can represent a distribution with strong modes $\mathcal{C}$.
- And, if $\text{RBM}_{n,m}$ can represent a distribution with strong modes $\mathcal{C}$, then $\mathcal{C}$ is a LTC.

- Combining these results gives our answer to

Problem

Given some $n, m \in \mathbb{N}$, what is the smallest $k \in \mathbb{N}$ for which the k-mixture of product distributions on n binary variables contains the RBM model with n visible and m hidden units?

- Namely, if $4\lceil m/3 \rceil \leq n$, then $\mathcal{M}_{n,k} \supseteq \mathrm{RBM}_{n,m}$ if and only if $k \geq 2^m$;
- and if $4\lceil m/3 \rceil > n$, then $\mathcal{M}_{n,k} \supseteq \mathrm{RBM}_{n,m}$ only if $k \geq \min\{2^l + m - l, 2^{n-1}\}$, where l is $\max\{l \in \mathbb{N} \colon 4\lceil l/3 \rceil \leq n\}$.
- Thus an exponentially larger mixture model, with an exponentially larger number of parameters, is required to represent distributions that can be represented by the RBM.

There's another way to see that the RBM has points of rank 2^m when $2m \leq n$

Not very often

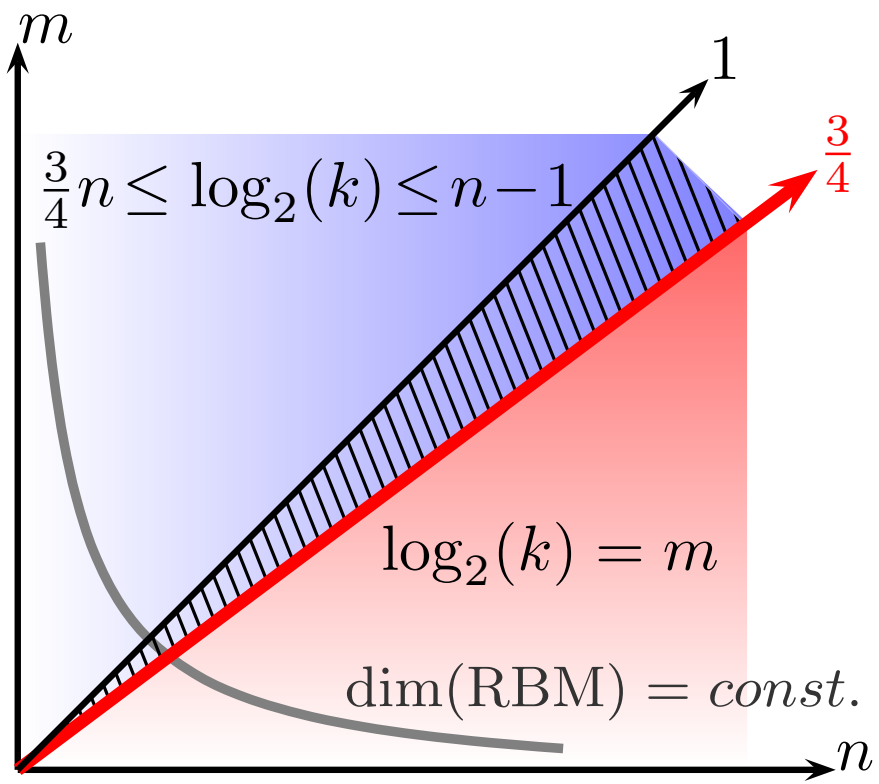


Figure: Plot of the smallest k for which $\mathcal{M}_{n,k}$ contains $\text{RBM}_{n,m}$. Fixing dimension (grey line), the RBMs which are hardest to represent as mixtures are those where $m/n \approx 1$.

1. Algebraic geometry of tensor networks
   - Tensors
   - Tensor Networks
   - Algebraic geometry

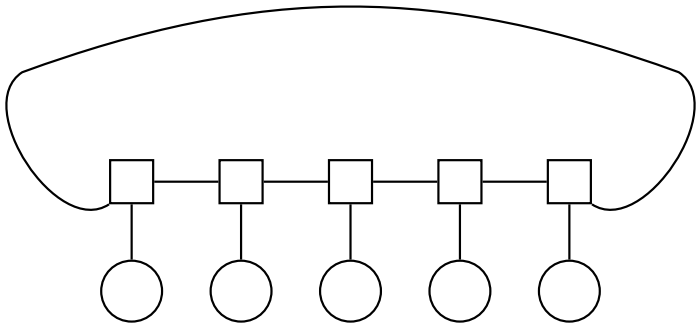
2. Algebraic Description of Graphical Models
   - Review of GM Defininitions
   - Algebraic and semialgebraic descriptions
   - Restricted Boltzmann machines

3. Identifiability, singular learning theory, other perspectives
   - Identifiability
   - Singular Learning Theory

Identifiability: uniqueness of parameter estimates

- A parameterization of a set of probability distributions is identifiable if it is injective.
- A parameterization of a set of probability distributions is generically identifiable if it is injective except on a proper algebraic subvariety of parameter space.
- Identifiability questions can be answered with algebraic geometry (e.g. many recent results in phylogenetics)
- A weaker question: What conditions guarantee generic identifiability up to known symmetries?
- A still weaker question: is the dimension of the space of representable distributions (states) equal to the expected dimension (number of parameters)? Or are parameters wasted?

Uniqueness up to known symmetries and normal forms



- Identify internal symmetries (here SL_2)
- Reparameterize to choose a normal form

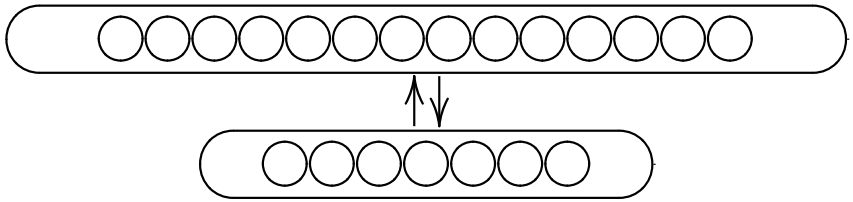
Singular learning theory

A model is more than its implicitization; the parameterization map is critically important to learning performance and quality.

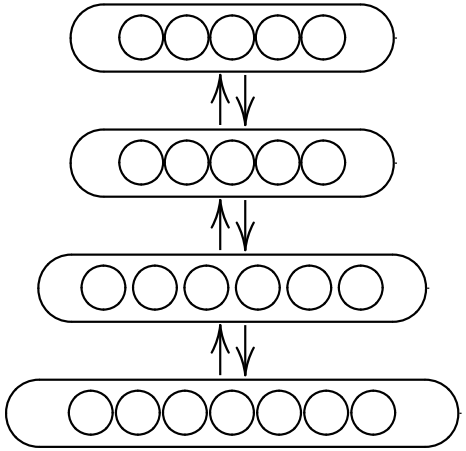
How fast and how well can a model learn?

- When a statistical model is regular, we can use central limit theorems to figure out their behavior for large data.
- But most hidden variable models are not regular (identifiable w/ positive definite Fisher information matrix) but singular.
- Singular learning theory [Watanabe 2009] offers one avenue for progress in this situation based on algebraic geometry.
- Asymptotics, generalization error, etc are governed by the real log canonical threshold.
- Resolve the model singularities and develop new limit theorems.

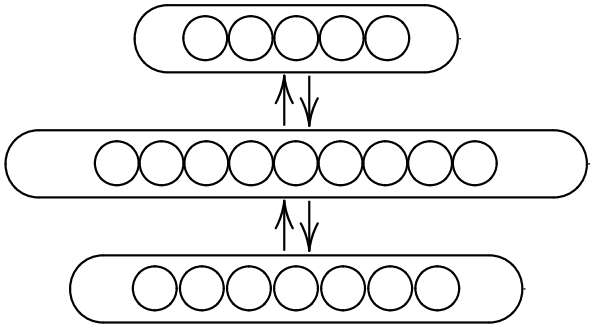
Comparing Architectures



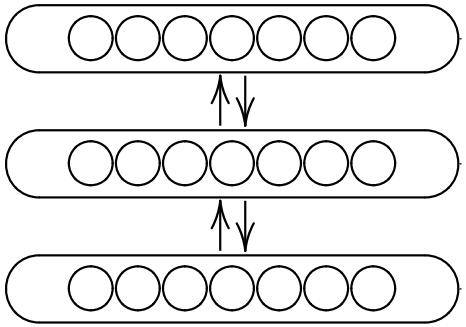
RBM
(9,18): 189 parameters



Treelike
(9,8,6,6): 185 parameters



Bulge
(9,11,6): 191 parameters



Column
(9,9,9): 189 parameters

Optimal architectures for learning

- The real log canonical threshold λ_q of a parameterization at true distribution $q = p(x|\theta)$ determines Bayes generalization error, $G_n(q) = \mathbb{E}_q[KL(q||p_n^*)] - S_q = \frac{\lambda_q}{n} + o(\frac{1}{n})$ [Watanabe 2009].
  - Expected KL-divergence
  - from the true model to the predicted distribution p^*
  - after seeing n observations and updating to the posterior.
- E.g. what do we have to believe about which qs appear in nature for the deep model to be better, $\lambda_{RBM}(q) > \lambda_{COL}(q)$?
- Techniques for calculating λ are rapidly evolving; known for simple binary graphical models such as trees [Zwiernik 2011].

Advertisement

- Modern applications of representation theory
  - an IMA PI Graduate Summer School
  - at the University of Chicago
  - Summer 2014
- $\approx$ 12 Lectures on tensor networks