SINGULAR LEARNING THEORY

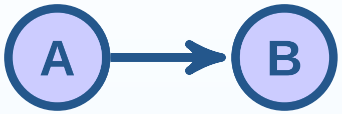
Part III: Singularities in Graphical Models

Shaowei Lin
(Institute for Infocomm Research, Singapore)

21-25 May 2013
Motivic Invariants and Singularities Thematic Program
Notre Dame University

Graphical Models

Basic Ingredients



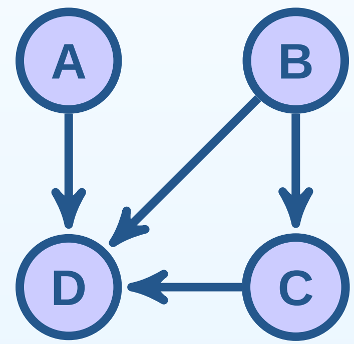
Directed edges: A causes B.

Undirected edges: A and B are correlated.

Graphical models are defined by a collection of random variables

e.g. $X = (X_A, X_B, X_C, X_D)$

and a graph $G = (V, E)$ describing the relationship between the variables.

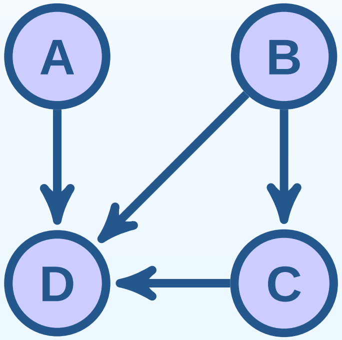


	Discrete	**Gaussian**
Directed Acyclic Graphs	Also known as Baysian networks.	
Undirected Graphs	Also known as Markov random fields.	

Directed Graphical Models

Factorization Property (parametric)

$$\mathbb{P}(X) = \prod_{v \in V} \mathbb{P}(X_v | X_{\text{parents}(v)})$$



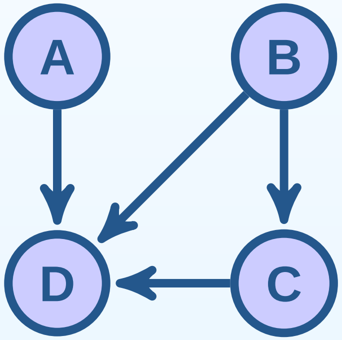
Discrete

$$\mathbb{P}(A, B, C, D) = \underbrace{\mathbb{P}(A)\,\mathbb{P}(B)}_{\text{root probabilities}}\,\underbrace{\mathbb{P}(C|B)\,\mathbb{P}(D|A, B, C)}_{\text{conditional probabilities}}$$

Directed Graphical Models

Factorization Property (parametric)

$$\mathbb{P}(X) = \prod_{v \in V} \mathbb{P}(X_v | X_{\text{parents}(v)})$$



Gaussian

$$A = \varepsilon_A, \qquad \varepsilon_A, \varepsilon_B, \varepsilon_C, \varepsilon_D \sim \mathcal{N}(0,1)$$

$$B = \varepsilon_B$$

$$C = \lambda_{BC} B + \varepsilon_C$$

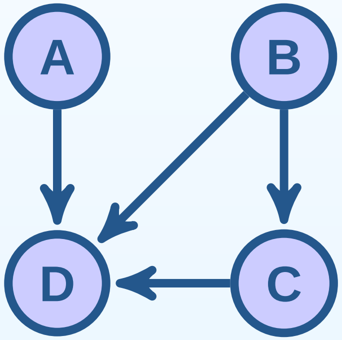
$$D = \lambda_{AD} A + \lambda_{BD} B + \lambda_{CD} C + \varepsilon_D$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\lambda_{BC} & 1 & 0 \\ -\lambda_{AD} & -\lambda_{BD} & -\lambda_{CD} & 1 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} \sim \mathcal{N}(0, \mathrm{Id})$$

Directed Graphical Models

Local Markov Property (implicit)

$$X_v \perp\!\!\!\perp X_{V \setminus \text{descendents}(v)} \mid X_{\text{parents}(v)} \quad \text{for all } v \in V$$



$$A \perp\!\!\!\perp B, C$$
$$B \perp\!\!\!\perp A$$
$$C \perp\!\!\!\perp A \mid B$$
$$D \perp\!\!\!\perp \emptyset \mid A, B, C$$

Global Markov Property (implicit)

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad \text{iff } A \text{ is d-separated from } B \text{ by } C$$

Directed Graphical Models

Hammersley-Clifford Theorem

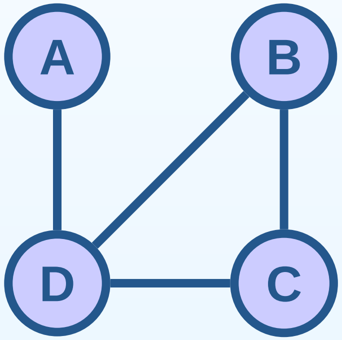
The following are equivalent:

- Factorization Property
- Local Markov Property
- Global Markov Property

Undirected Graphical Models

Factorization Property (parametric)

$$\mathbb{P}(X) = \frac{1}{Z} \prod_{\text{max-clique } C} \varphi_C(X_C), \quad Z \text{ normalizing const.}$$



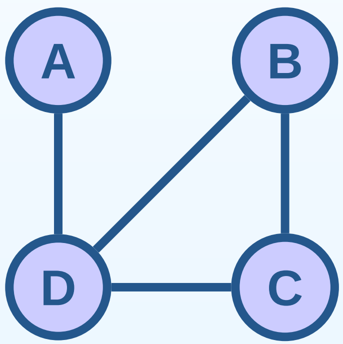
Discrete

$$\mathbb{P}(A,B,C,D) = \frac{1}{Z}\varphi_{AD}(A,D)\,\varphi_{BCD}(B,C,D)$$

Undirected Graphical Models

Factorization Property (parametric)

$$\mathbb{P}(X) = \frac{1}{Z} \prod_{\text{max-clique } C} \varphi_C(X_C), \quad Z \text{ normalizing const.}$$



Gaussian

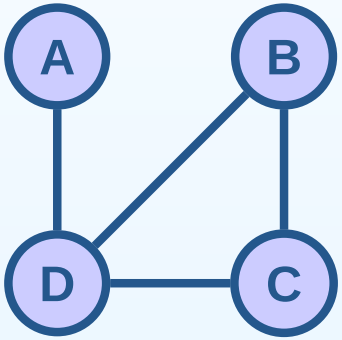
$$X = (X_v)_{v \in V} \sim \mathcal{N}(0, \Sigma)$$

such that $(\Sigma^{-1})_{uv} = 0$ iff $(u, v) \in E$.

Undirected Graphical Models

Local Markov Property (implicit)

$$X_v \perp\!\!\!\perp X_{\{v\} \cup \mathrm{nonneighbors}(v)} \mid X_{\mathrm{neighbors}(v)} \quad \text{for all } v \in V$$



$$A \perp\!\!\!\perp B, C \mid D$$
$$B \perp\!\!\!\perp A \mid C, D$$
$$C \perp\!\!\!\perp A \mid B, D$$
$$D \perp\!\!\!\perp \emptyset \mid A, B, C$$

Global Markov Property (implicit)

$$X_A \perp\!\!\!\perp X_B \mid X_C \quad \text{iff } A \text{ is separated from } B \text{ by } C$$

Directed Graphical Models

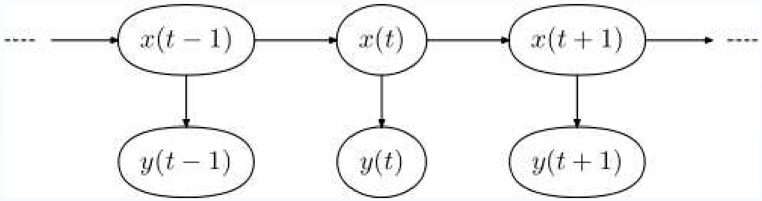
Hammersley-Clifford Theorem

If $\mathbb{P}(X = x) > 0$ for all x,
then the following are equivalent:

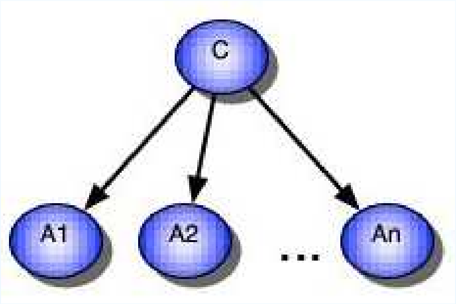
- Factorization Property
- Local Markov Property
- Global Markov Property

Important Graphical Models

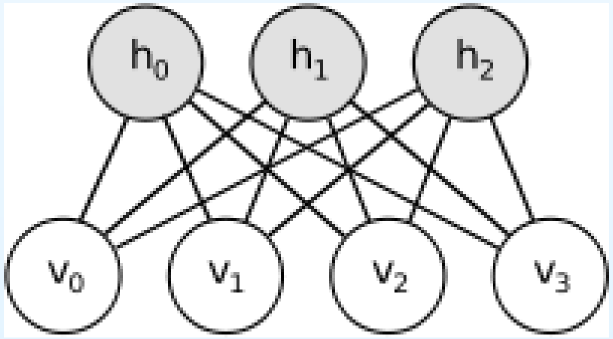
Hidden Markov Models



Gaussian Mixtures

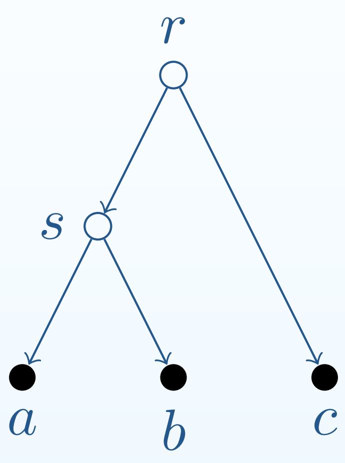


Restricted Boltzmann Machines



Tree Cumulants

Tree Models with Binary States



Parameters (9-dim): for each $(i,j) \in E$,

$$\pi_r = \mathbb{P}(X_r = 0),$$
$$t_j^{i0} = \mathbb{P}(X_j = 0 | X_i = 0),$$
$$t_j^{i1} = \mathbb{P}(X_j = 0 | X_i = 1).$$

Probabilities (7-dim): for each $I \subset \{a,b,c\}$,

$p_I = \mathbb{P}(X_i = 1$ for $i \in I, X_i = 0$ otherwise$)$.

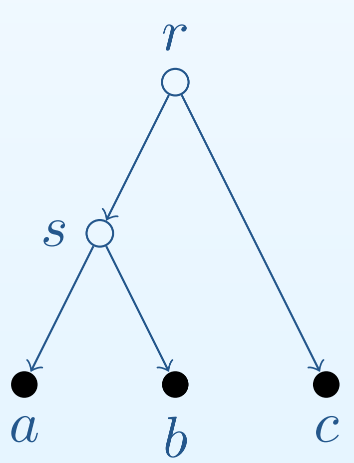
e.g.
$$\begin{aligned} p_{ab} = & \textstyle\sum_{i,j} \mathbb{P}_r(i)\mathbb{P}_{s|r}(j|i)\mathbb{P}_{a|s}(1|j)\mathbb{P}_{b|s}(1|j)\mathbb{P}_{c|r}(0|i) \\ = & \pi_r t_s^{r0}(1-t_a^{s0})(1-t_b^{s0})t_c^{r0} \\ & + \pi_r(1-t_s^{r0})(1-t_a^{s1})(1-t_b^{s1})t_c^{r0} \\ & + (1-\pi_r)t_s^{r1}(1-t_a^{s0})(1-t_b^{s0})t_c^{r1} \\ & + (1-\pi_r)(1-t_s^{r1})(1-t_a^{s1})(1-t_b^{s1})t_c^{r1} \end{aligned}$$

Compute the RLCT of the fiber ideal $\langle p_a - \hat{p}_a, \ldots, p_{abc} - \hat{p}_{abc} \rangle$?

Reparametrization

Strategy: Transform both the parameter space Ω and distribution space Δ so that the resulting map $\widetilde{\Omega} \to \widetilde{\Delta}$ is almost monomial.

$$\begin{array}{rccccl} \text{transitions} & \Omega & \longrightarrow & \Delta & \text{probabilities} \\ & \downarrow & & \downarrow & \\ \text{regressions} & \widetilde{\Omega} & \longrightarrow & \widetilde{\Delta} & \text{cumulants} \end{array}$$



Regressions (9-dim):

$\lambda_r, \lambda_s, \mu_a, \mu_b, \mu_c,$

$\eta^r_s, \eta^r_c, \eta^s_a, \eta^s_b.$

Cumulants (7-dim):

$k_a, k_b, k_c, k_{ab}, k_{bc}, k_{ac}, k_{abc}.$

Cumulant Equations

$$\begin{array}{rcccl} \text{transitions} & \Omega & \longrightarrow & \Delta & \text{probabilities} \\ & \downarrow & & \downarrow & \\ \text{regressions} & \widetilde{\Omega} & \longrightarrow & \tilde{\Delta} & \text{cumulants} \end{array}$$

$$k_a = \mu_a, \quad k_{bc} = \tfrac{1}{4}(1-\lambda_r^2)\eta_s^r\eta_b^s\eta_c^r,$$
$$k_b = \mu_b, \quad k_{ac} = \tfrac{1}{4}(1-\lambda_r^2)\eta_s^r\eta_a^s\eta_c^r,$$
$$k_c = \mu_c, \quad k_{ab} = \tfrac{1}{4}(1-\lambda_s^2)\eta_a^s\eta_b^s,$$
$$k_{abc} = \tfrac{1}{4}(1-\lambda_r^2)\lambda_s\eta_s^r\eta_a^s\eta_b^s\eta_c^r.$$

Cumulants recently extended to non-binary non-tree models.

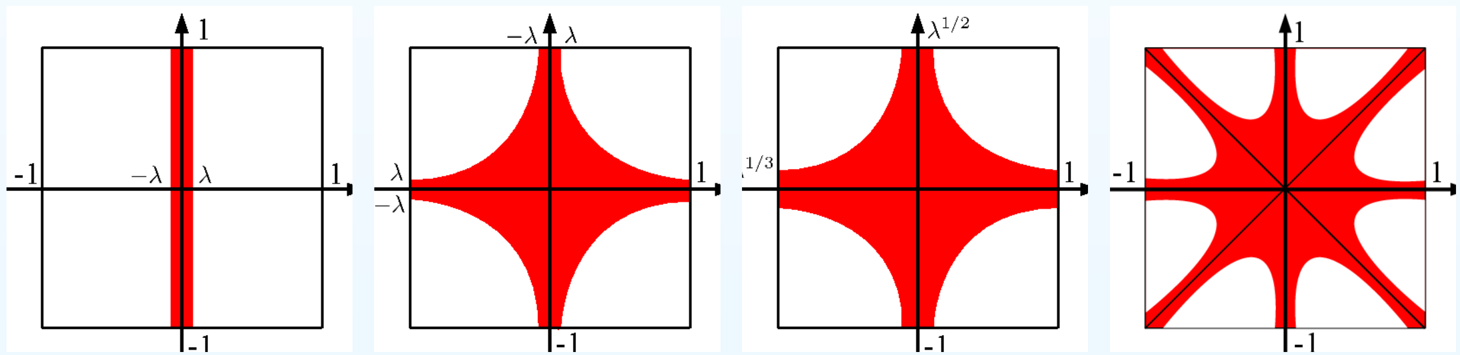
Statistics give new insights to difficult algebraic geometry problems.

- J. Q. Smith, P. Zwiernik: Tree-cumulants and the geometry of binary tree models, Bernoulli **18** (2012), 290–321.
- P. Zwiernik: An Asymptotic Behaviour of the Marginal Likelihood for General Markov Models, J. of Machine Learning Research **12** (2011), 3283–3310.
- B. Sturmfels, P. Zwiernik: Binary cumulant varieties, Ann. Combinatorics **17** (2013), 229–250.
- M. Michałek, L. Oeding, P. Zwiernik: Secant cumulants and toric geometry, arXiv:1212.1515.

Partial Correlation Algorithm

Volumes of Tubular Neighborhoods

Real log canonical thresholds also allow us to approximate the volume of small tubular neighborhoods of varieties. Such problems occur frequently in error estimation and convergence analysis.



(a) x (b) xy (c) x^2y^3 (d) $x^3y - xy^3$

Tubes $|f(x,y)| \le t$ for various polynomials in two variables.

Partial Correlation Algorithm

Partial Correlations.

Gaussian model with variables V, concentration matrix $K = \Sigma^{-1}$.
Given $i, j \in V$ and $S \subset V \setminus \{i, j\}$, let $R = V \setminus (S \cup \{i, j\})$.

$$\mathrm{corr}_{i,j|S} = \frac{\det(K_{iR,jR})}{\sqrt{\det(K_{iR,iR}) \cdot \det(K_{jR,jR})}}$$

The partial correlation (PC) algorithm constructs directed Gaussian graphical models by inferring conditional independence statements $i \perp\!\!\!\perp j \mid S$ from the data.

1. Fix a small tolerance $t > 0$.
2. Start with a complete graph G.
3. Run through all triples (i, j, S), $i, j \notin S$, systematically.
4. For each (i, j, S), compute the partial correlation $\mathrm{corr}_{i,j|S}$.
5. If $\mathrm{corr}_{i,j|S} \leq t$, then remove edge (i, j) from graph G.

Faithfulness

- A distribution $p(\cdot|\omega)$ is t-strong-faithful to a graph G if

  $|\mathrm{corr}_{i,j|S}(\omega)| \leq t \;\Leftrightarrow\; i$ is d-separated from j given S.

  Otherwise, it is unfaithful.

- Using singular learning theory, we can approximate the volume of unfaithful parameters as t goes to zero.

$$\int_{|f(\omega)|\leq t} d\omega \;\approx\; Ct^{\lambda}(-\log t)^{\theta-1}$$

  Here (λ, θ) is the learning coefficient of $f(\omega) = \mathrm{corr}_{i,j|S}(\omega)$. Determines performance of the PC algorithm for large samples.

- e.g. $(\lambda, \theta) = \begin{cases} (1,1) & \text{for all star trees,} \\ (1, p-1) & \text{for a chain with } p \text{ nodes.} \end{cases}$

- S. Lin, C. Uhler, B. Sturmfels, and P. Bühlmann: Hypersurfaces and their singularities in partial correlation testing, arXiv:1209.0285.

Sparse Model Selection

The Big Data Phenomenon

Examples of applications

- image recognition
- speech recognition
- language translation
- sentiment analysis
- pedestrian detection
- bioinformatics
- healthcare planning
- recommendation systems

Characteristics of Big Data

1. High dimensional data vectors
2. Data cuts out a low dimensional manifold
3. Learning a model with high dimensional parameter space
4. Very large sample sizes

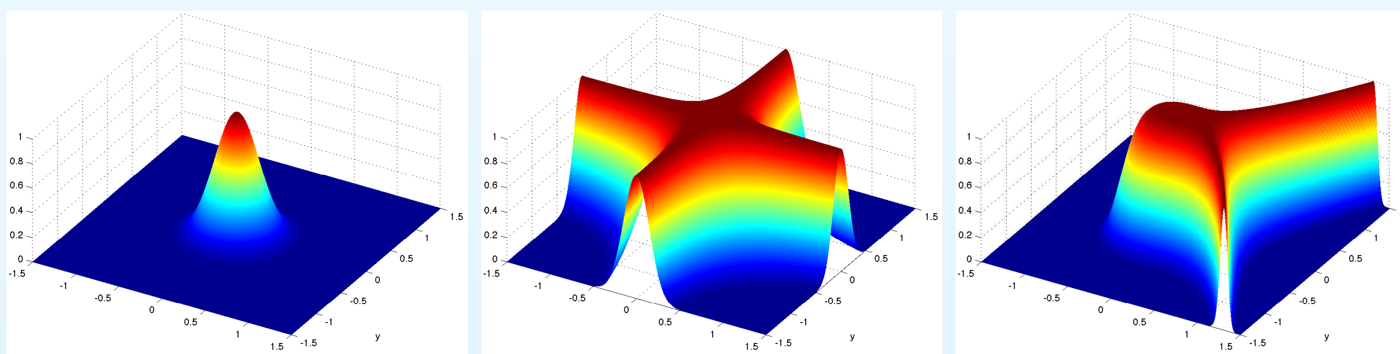
Curse of Singularities

- Approximation of high dimensional integrals is difficult because of the curse of ~~dimensionality~~ singularities.

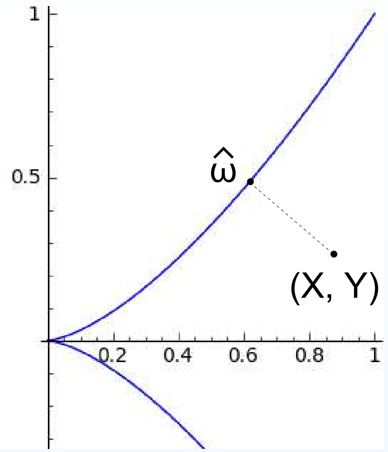
  For smooth models, Laplace approximation works well even if parameter space $\mathbb{R}^d$ has high dimension.

- But many models in machine learning are singular, e.g. mixtures, neural networks, hidden variables.

- Important to analyze asymptotics of integrals with singularities.



Learning a Singular Model



$$X \sim \mathcal{N}(\omega^2, 1), \quad Y \sim \mathcal{N}(\omega^3, 1)$$

data $(X_i, Y_i), i = 1 \,..\, N$

parameter $\omega \in \mathbb{R}$, mean $(\bar{X}, \bar{Y})$

- MLE: $\operatorname{argmin}_\omega |\omega^2 - \bar{X}|^2 + |\omega^3 - \bar{Y}|^2$
  BIC performs poorly when MLE is close to 0.

- Recall that the likelihood integral is

$$Z_N = \frac{1}{2\pi} \int_{\mathbb{R}} \exp(-\frac{1}{2} \sum_{i=1}^{N} |\omega^2 - X_i|^2 + |\omega^3 - Y_i|^2)\, d\omega$$

- If true distribution is $X \sim \mathcal{N}(u^2, 1), Y \sim \mathcal{N}(u^3, 1)$, then

$$-\log Z_N(u) \approx \frac{1}{2} \sum_{i=1}^{N} (u^2 - X_i)^2 + (u^3 - Y_i)^2 + \pi(u) + O_p(1)$$

where $\pi(u) = \frac{1}{4} \log N$ if $u = 0$; otherwise $\pi(u) = \frac{1}{2} \log N$.

Learning Coefficients

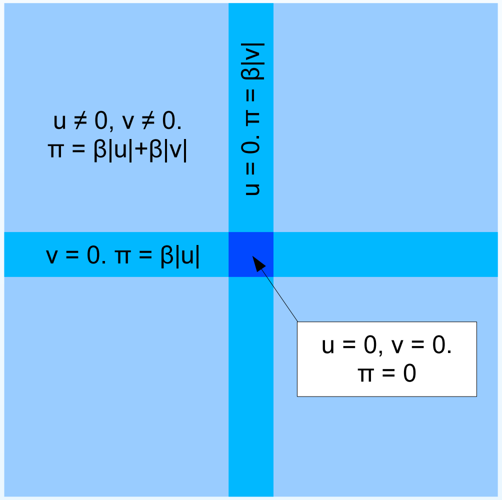
- Given $u \in \Omega$, there exist learning coefficients (λ_u, θ_u) such that for all sufficiently small nbhds Ω_u of u,
$$\int_{\Omega_u} e^{-Nf(\omega)}\,\varphi(\omega)d\omega \;\approx\; CN^{-\lambda_u}(\log N)^{\theta_u - 1}.$$
- Sparsity penalty for MLE: Given the log likelihood
$$\ell(u) = -\textstyle\sum_{i=1}^N \log p(X_i|u),$$
for large samples we have the asymptotic approximation
$$-\log Z(u) \;\approx\; \ell(u) + \pi(u) + O_p(1)$$
where $\pi(u) = \lambda_u \log N - (\theta_u - 1)\log\log N$.
- To find the model $\mathcal{M}_u$ that minimizes $Z(u)$, we compute
$$\operatorname{argmin}_{u\in\Omega} l(u) + \pi(u).$$
- This is a generalization of the BIC to singular models.

Computational Problems

- How do we generalize compressive sensing to singular models?

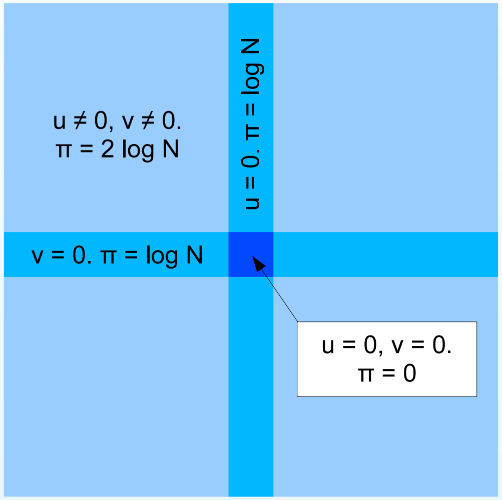
Compressive Sensing

$$\pi(\omega) = |\omega|_1 \cdot \beta$$



Bayesian Info Criterion (BIC)

$$\pi(\omega) = |\omega|_0 \cdot \log N$$



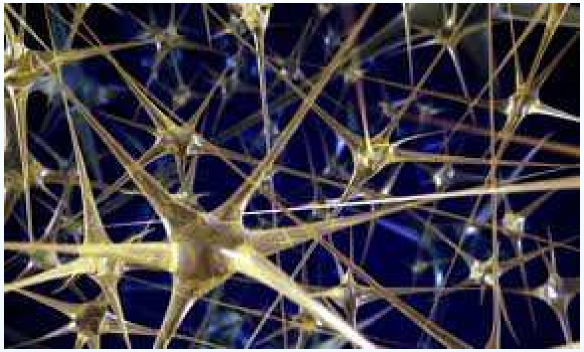
(Parameter space partitioned into regions with different weights.)

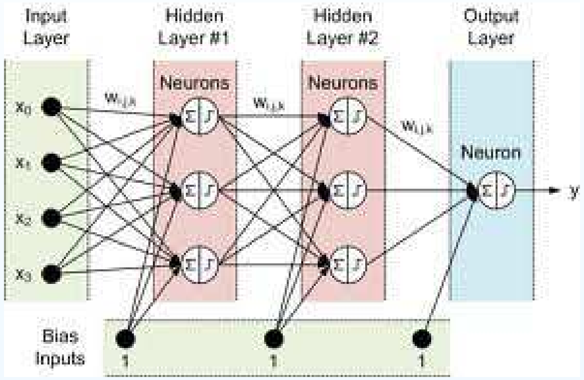
- How do we use RLCTs to improve MCMC techniques?

Neural Networks

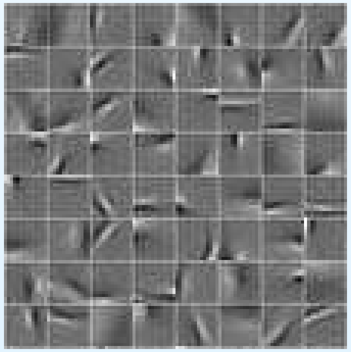
Motivation for Neural Networks

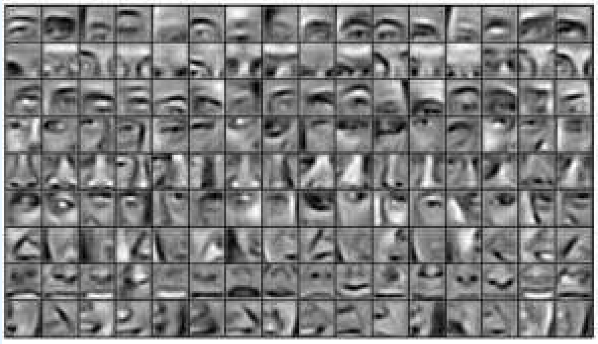
Neural networks are highly singular models inspired by biology.





The lack of success forced researchers to abandon these models in the 70's and 80's. But the introduction of multiple layers and nonlinear sparse methods turned the tide. Computationally fast.





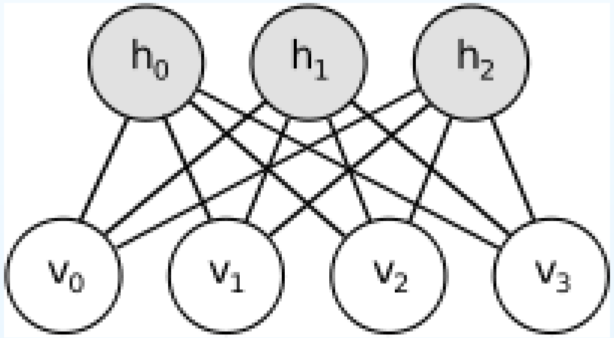
Singular learning tells us that proper learning requires sparsity.

Restricted Boltzmann Machines

Undirected discrete graphical model with binary states.

Graph G is bipartite with two layers: observed and hidden nodes.

Parameters: weights ω_{ij} for each edge (h_i, v_j),
bias b_i for each h_i, bias c_j for each v_j.



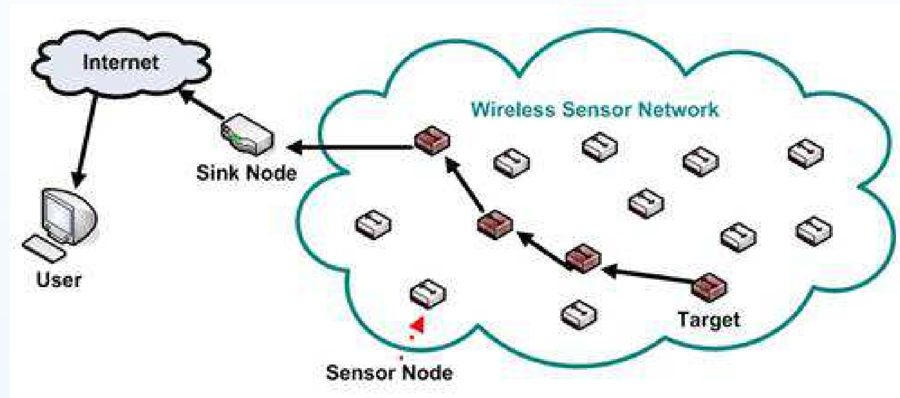
$$\mathbb{P}(h_i = 1|v) = \mathrm{sig}(b_i + \sum_j \omega_{ij} v_j), \quad \mathrm{sig}(x) = \frac{1}{1+e^{-x}}$$

Mimics behavior of biological neurons! Used in Deep Learning.

Tropical geometry used to find the dimension of the RBM.
Some RLCTs were also computed (via desingularizations).

Wireless Sensor Networks

Wireless network of sensors communicating real-time information.



Machine learning principles (graphical models, sparsity, singularities) for analyzing and designing wireless sensor networks (data compression, transmission, network connectivity, security).

sense & sense abilities

WE ARE HIRING!

Bachelors/PhD with strong background in mathematics and machine learning.

Thank you!

"Algebraic Methods for Evaluating Integrals in Bayesian Statistics"

`http://math.berkeley.edu/~shaowei/swthesis.pdf`

(PhD dissertation, May 2011)

References

1. V. I. Arnol'd, S. M. Guseĭn-Zade and A. N. Varchenko: Singularities of Differentiable Maps, Vol. II, Birkhäuser, Boston, 1985.
2. M. Aoyagi: Stochastic Complexity and Generalization Error of a Restricted Boltzmann Machine in Bayesian Estimation, J. Mach. Learn. Res. **99** (2010) 1243–1272.
3. A. Bravo, S. Encinas and O. Villamayor: A simplified proof of desingularisation and applications, Rev. Math. Iberoamericana **21** (2005) 349–458.
4. D. A. Cox, J. B. Little, and D. O'Shea: Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra, Springer-Verlag, New York, 1997.
5. M. Cueto, J. Morton and B. Sturmfels: Geometry of the Restricted Boltzmann Machine, Algebraic Methods in Statistics and Probability II, Contemporary Mathematics, AMS **516** (2010), 135–153.
6. M. Evans, Z. Gilula and I. Guttman: Latent class analysis of two-way contingency tables by Bayesian methods, Biometrika **76** (1989) 557–563.
7. H. Hironaka: Resolution of singularities of an algebraic variety over a field of characteristic zero I, II, Ann. of Math. (2) **79** (1964) 109–203.
8. S. Lin, B. Sturmfels and Z. Xu: Marginal likelihood integrals for mixtures of independence models, J. Mach. Learn. Res. **10** (2009) 1611–1631.
9. S. Lin: Algebraic methods for evaluating integrals in Bayesian statistics, PhD dissertation, Dept. Mathematics, UC Berkeley (2011).
10. S. Watanabe: Algebraic Geometry and Statistical Learning Theory, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, Cambridge, 2009.