The Pleasures of Counting

DLD and Jared Tanner

Tanner: University of Edinburgh

May 17, 2009

Reminiscences

- *Clustering Algorithms*
- Harvard-Yale days
- John's opinions

Some papers I lived inside

- Uniform Convergence of the Empirical Distribution Function Over Convex Sets W. F. Eddy, J. A. Hartigan The Annals of Statistics, Vol. 5, No. 2 (Mar., 1977), pp. 370-374
- Representing Points in Many Dimensions by Trees and Castles B. Kleiner, J. A. Hartigan Journal of the American Statistical Association, Vol. 76, No. 374 (Jun., 1981), pp. 260-269
- The Dip Test of Unimodality J. A. Hartigan, P. M. Hartigan The Annals of Statistics, Vol. 13, No. 1 (Mar., 1985), pp. 70-84

Springer Series in Statistics

L. A. Goodman and W. H. Kruskal, Measures of Association for Cross Classifications. x, 146 pages, 1979.

J. O. Berger, Statistical Decision Theory: Foundations, Concepts, and Methods. xiv, 425 pages, 1980.

R. G. Miller, Jr., Simultaneous Statistical Inference, 2nd edition. xvi, 299 pages, 1981.

P. Brémaud, Point Processes and Queues: Martingale Dynamics. xviii, 354 pages, 1981.

E. Seneta, Non-Negative Matrices and Markov Chains. xv, 279 pages, 1981.

F. J. Anscombe, Computing in Statistical Science through APL. xvi, 426 pages, 1981.

J. W. Pratt and J. D. Gibbons, Concepts of Nonparametric Theory. xvi, 462 pages, 1981.

V. Vapnik, Estimation of Dependences based on Empirical Data. xvi, 399 pages, 1982.

H. Heyer, Theory of Statistical Experiments. x, 289 pages, 1982.

L. Sachs, Applied Statistics: A Handbook of Techniques. xxviii, 706 pages, 1982.

M. R. Leadbetter, G. Lindgren and H. Rootzen, Extremes and Related Properties of Random Sequences and Processes. xii, 336 pages, 1983.

H. Kres, Statistical Tables for Multivariate Analysis. xxii, 504 pages, 1983.

J. A. Hartigan, Bayes Theory. xii, 145 pages, 1983.

J. A. Hartigan

Bayes Theory

Springer-Verlag
New York Berlin Heidelberg Tokyo

CHAPTER 9

9.5. When Most of the Means Are Small

In 9.2 and 9.3, g is normal with mean 0 and unknown variance, and a prior distribution is placed on the variance. In many regression and analysis of variance problems, most of the means X_i are very close to zero, but a few are quite large. Such a situation is not well represented by a normal g, because it is not sufficiently long tailed. One alternative is to assume that X_i comes from a distribution $p\delta_0 + (1-p)N(0, \sigma_0^2)$ where $\delta_0\{0\} = 1$. Then Y_i is a random sample from $pN(0, 1) + (1-p)N(0, \sigma_0^2 + 1)$,

$$g(y) = \frac{1}{\sqrt{2\pi}}\left\{p\exp(-\tfrac{1}{2}y^2) + \frac{(1-p)}{\sqrt{1+\sigma_0^2}}\exp(-\tfrac{1}{2}y^2/(1+\sigma_0^2))\right\}$$

$$\frac{d}{dy}\log g(y) = \frac{-y\left\{p\exp(-\tfrac{1}{2}y^2) + \dfrac{(1-p)}{(1+\sigma_0^2)^{3/2}}\exp(-\tfrac{1}{2}y^2/(1+\sigma_0^2))\right\}}{\left\{p\exp(-\tfrac{1}{2}y^2) + \dfrac{1-p}{\sqrt{1+\sigma_0^2}}\exp(-\tfrac{1}{2}y^2/(1+\sigma_0^2))\right\}}$$

If y is small, the adjustment is close to $-y$; if y is large it is close to $y/(1+\sigma_0^2)$; in this way the small observed values Y_i are moved very close to zero, but the large observed values Y_i are relatively unchanged. In practice p and σ_0^2 must be estimated from the Y_i. A Bayesian approach requires computation of the posterior mean of $(d/dy)\log g(y)$ but no prior on p and σ_0 is known which permits explicit computation. It is straightforward computationally to estimate p and σ_0^2 to maximize the likelihood of the observations, but explicit expressions are not available, and it is not known whether the resulting estimates of the X_i beat the straight estimates.

A standard approach to the problem of many small means is to carry out a significance test on each mean separately, and to set all those means to zero which do not exceed some significance level. Here, the estimate would be $\hat{X}_i = Y_i\{|Y_i| \geqq c\}$, where c is the cutoff point in the significance test. Then

$$\sum P(Y_i - X_i)^2 - \sum P(\hat{X}_i - X_i)^2 = \sum P(\{|Y_i| < c\}(Y_i^2 - 2Y_iX_i))$$
$$= \sum P(\{|Z_i + X_i| < c\}(Z_i^2 - X_i^2))$$

where $Z_i \sim N(0, 1)$. If $|X| > 1$, $P(\{|Z + X| < c\}(Z^2 - X^2)) < 0$ for every choice of c. Thus there is no way to choose c so that the estimates $\hat{X}_i$ have uniformly smaller mean square error than Y_i; it does not help to allow c to depend on the Y_i.

Yet there is practical value in setting many small means to be exactly zero if there is no evidence of significant departure from zero. Suppose the loss function is

$$L(d, s) = \{d \neq s\} + k(d - s)^2.$$

Let P_0 be a unitary probability on S which has an atom $P_0\{s_0\}$ only at s_0. Then the probable loss for d is $P_0\{d \neq s\} + kP_0(d - s)^2 = P_0 s_0 + \{d = s_0\}(1 - 2P_0 s_0) + k(d - P_0 s)^2 + kP_0(s - P_0 s)^2$. The Bayes decision is $d = s_0$ if $2P_0\{s_0\} > k(s_0 - P_0 s)^2 + 1$, and $d = P_0 s$ otherwise.

A FAILURE OF LIKELIHOOD ASYMPTOTICS FOR NORMAL MIXTURES

J. A. HARTIGAN Yale University

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_n$ be a sample from the normal mixture

$$(1 - p)N(0, 1) + pN(\theta, 1).$$

To test $\theta = 0$ against $\theta \neq 0$, we might use the likelihood ratio test based on

$$L_n = \sup_{\theta, p} L_n(\theta, p),$$

where $L_n(\theta, p) = \sum_{i=1}^{n} \ln[(1 - p) + p \exp(X_i\theta - \frac{1}{2}\theta^2)]$. Tests of this type are used in clustering to decide how many clusters or components in the normal mixture model are present. (More generally, the normal mixture contains k multivariate normal components with means and covariance matrices different for each component.)

When $\theta = 0$, it makes no difference what value p has; thus, $L_n(\theta, p)$ is maximized over two parameters, and $L_n(0, p) = 0$; standard asymptotics has L_n asymptotically $\frac{1}{2}\chi_2^2$.

From *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Volume II, Lucien M. Le Cam and Richard A. Olshen, eds.,

Back to the Future

- Today: Genomics, Proteomics, Metabolonomics, Medical Imaging
  - Features $X_{i,j}, j = 1, \dots, p$ of gene expression, or spectrum, or ...
  - Response Y_i: Categorical (e.g. disease $\pm$) or Continuous (e.g. survival time)
- Large p, Fixed n
  - $p \geq 10K$ (Gene expression, SNP, Mass Spectra, ...)
  - $n \leq 1K$ (Affecteds, Compliants,...)
- Goals:
  - Prediction: $Y = X\beta + Z$, eg Z normal
  - Classification: $E(X|\pm) = \beta_0 \pm \beta$
- *Generalization of Many Normal Means!*

The Hydrogen Atom: Underdetermined Linear Systems

- Alternate notation, *rest of talk*
  - A an $n \times N$ matrix, $n < N$.
  - x_0 unknown vector
  - $y = Ax_0$.
- Correspondence to Statistics
  - $Y \leftrightarrow y$
  - $X \leftrightarrow A$
  - $\beta \leftrightarrow x_0$
  - $n \leftrightarrow n$
  - $p \leftrightarrow N$
  - $p > n \leftrightarrow N > n$.
- Problem: the system $y = Ax$ is underdetermined; infinitely many solutions, no hope to determine x_0.
- What if most of the entries in x_0 are zero or nearly zero?

ℓ_1-ℓ_0 equivalence

- Sparsest Solution

$$(P_0) \qquad \min \|x\|_0 \text{ subject to } y = Ax$$

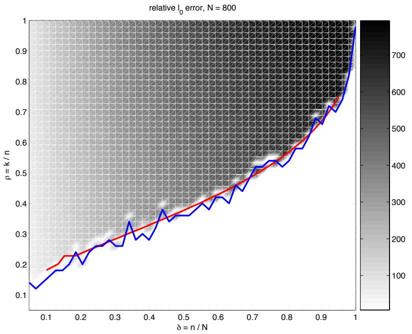
- NP Hard in general
- Minimal ℓ_1 Solution

$$(P_1) \qquad \min \|x\|_1 \text{ subject to } y = Ax$$

- In many cases the two problems have the same, unique, solution.
- Much literature on this topic; IEEE IT 2001-today

Empirical Results

Gaussian Random Matrix A, $\delta = k/n$, $\rho = n/N$.



$\rho_W^{\pm}(\delta)$: DLD (2004,2006) DLD and Jared Tanner, JAMS 2008

Nonnegative coefficients

- Underdetermined system of equations:

$$y = Ax, \qquad x \geq 0$$

- Sparsest Solution

$$(NP) \qquad x_0 = \text{argmin}\,\|x\|_0 \text{ s.t. } y = Ax, \quad x \geq 0$$

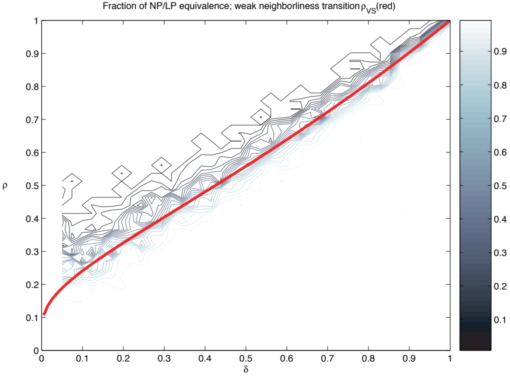
- Problem: NP-hard in general.
- Relaxation:

$$(LP) \qquad x_1 = \text{argmin}\,1'x \text{ s.t. } y = Ax, \quad x \geq 0.$$

Convex optimization – linear program

Empirical Results

Jared Tanner (2005)



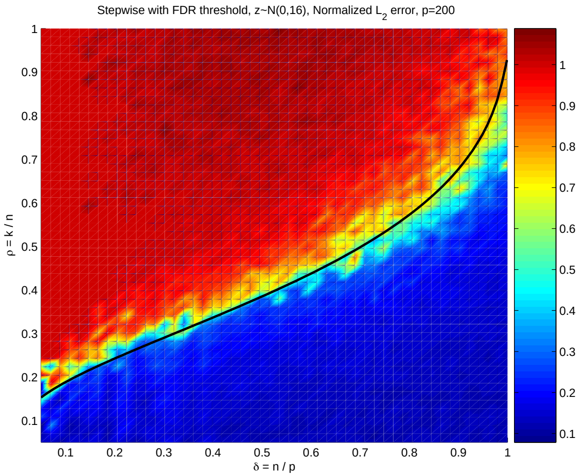
$\rho_W^+(\delta)$: DLD and Jared Tanner, PNAS (2005)

Appearance in Stepwise Regression

Victoria Stodden, Stanford Thesis, 2006

- $p = 200$ potential predictors
- n observations, a fraction $\delta = n/p \in (0, 1)$ of number of potential predictors
- k *useful* predictors, a fraction $\rho = k/n \in (0, 1)$ of number of observations
- Gaussian white noise
- $U(0, 1)$ regression coefficients
- Forward Stepwise Regression
- False Discovery Rate Stopping rule ($FDR \leq \frac{1}{2}$).

Phase Diagram for Stepwise FDR



Appearance in Computational Complexity

- Minimal ℓ_1 Solution

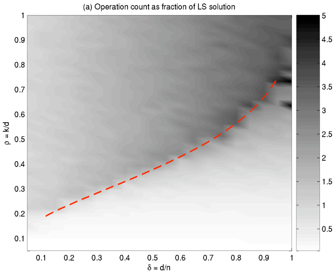
$$(P_1) \qquad \min \|x\|_1 \text{ subject to } y = Ax$$

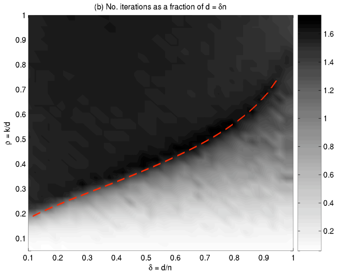
- Mixed ℓ_1-ℓ_2 objective

$$(H_\lambda) \qquad \min \|y - Ax\|_2^2/2 + \lambda\|x\|_1$$

- As $\lambda \rightarrow 0$ solves ℓ_1.
- Solution Path (x_λ) is polygonal.
- Osborne et al. (1999)
- Efron, Hastie Johnstone, Tibshirani (2003)

Solution Cost of ℓ_1 minimization problems





DLD and Tsaig IEEE IT 2008

Take-Away Messages for Today

- Phase Diagrams fundamental tool
- Phase Transitions for key observables in large N.
- Rigorous theory for Gaussian Random A
  Geometric Combinatorics.
- Conclusive Evidence:
  - Universality across broad classes of random matrices A
  - Extends to other algorithms
  - Extends to noisy observations
- Applications: Compressed Sensing

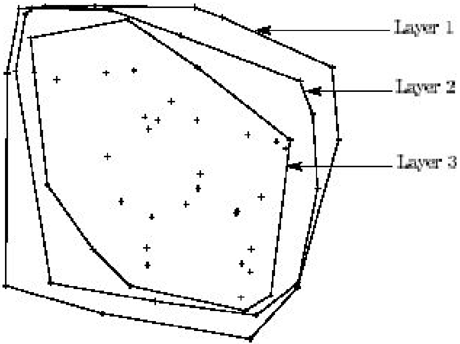
Random Convex Hulls

- X_i n random points in $\mathbf{R^d}$
- X_i, $i = 1, \dots, n$ iid $N(0, \Sigma)$
- $P = P_{d,n} = conv\{X_i\}$
- Classically: d fixed

$$\#vertP \sim c_d \log^{(d-1)/2} n, \quad n \to \infty.$$

- Renyi-Sulanke (1963), Efron (1965), Raynaud (1971), Hueter (1998)

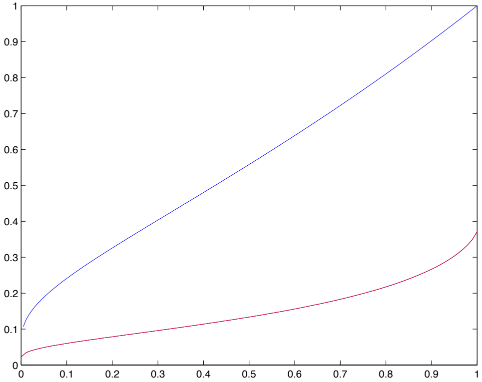
Our Low-Dimensional Intuition



Modern Era: High-Dimensional Case

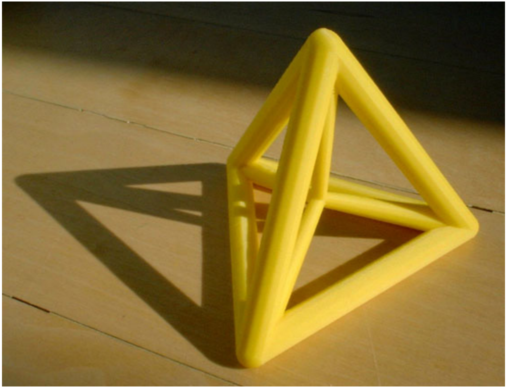
- d proportional to n, both large.
- n larger than d but only proportionally so
- **Surprise:**
  - *Every* X_i is a vertex
  - *Every* (X_i, X_j) span an edge
  - *Every* k-tuple spans a $k-1$ face, $k = 1, 2, \dots, \rho d$
  - ... up to quite large k!
- $\rho = k/d$
- $\delta = d/n$

Weak/Strong Transition $\rho_W^+(\delta)$,$\rho_S^+(\delta)$

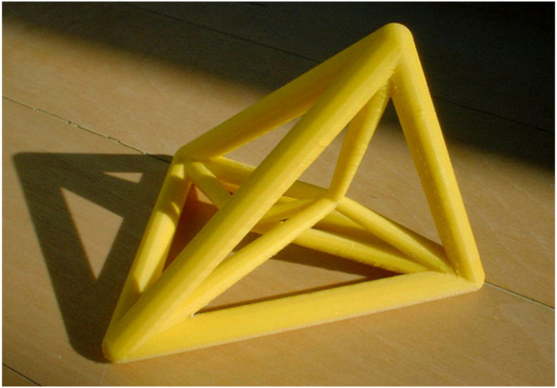


Donoho, Tanner, PNAS (2005)

4 Simplex Projected to $\mathbf{R}^3$



5 Simplex Projected to $\mathbf{R}^3$



Symmetrized Gaussian Point Cloud

- X_i n random points in $\mathbf{R^d}$
- X_i, $i = 1, \dots, n$ iid $N(0, \Sigma)$
- $P_\pm = conv\{\pm X_i\}$
- Classically: d fixed

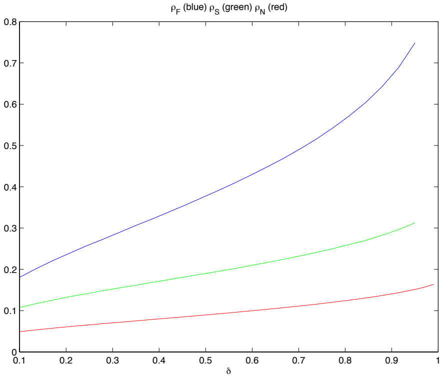
$$\#vertP \sim c'_d \log^{(d-1)/2} n, \quad n \to \infty.$$

- Böroczky and Henk (1998)

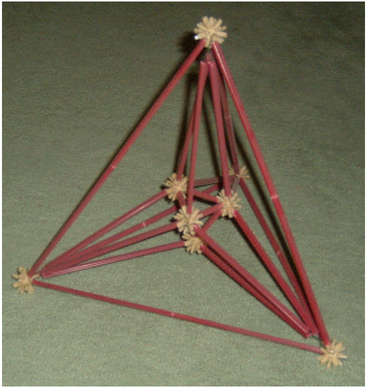
High-Dimensional Case

- d proportional to n!
- Novel Setting:
  - *Every* $\pm X_i$ is a vertex
  - *Every* $(\pm X_i, \pm X_j)$ span an edge
  - *Every* k-tuple spans a $k-1$ face, $k = 1, 2, \ldots, \rho d$, provided it contains no antipodal pair.
  - This continues up to some threshold in k.
- $\rho = k/d$
- $\delta = d/n$

Strong/Weak Transitions $\rho_S^{\pm}(\delta)$



4D Cross-Polytope Projected to $\mathbf{R}^3$



Correspondence: Faces $\leftrightarrow$ Models

- Collection of statistical models: poset.

$$\mathcal{M}_0 = \{X_3, X_{12}, X_{106}\}, \qquad \mathcal{M}_1 = \{X_3, X_{12}, X_{21}, X_{106}, X_{203}\}.$$

- Faces of a polytope: poset

$$F_0 = \{x : x \geq 0, x_j = 0, j \notin \{3, 12, 106\}, \sum_j x_j = 1\},$$

$$F_1 = \{x : x \geq 0, x_j = 0, j \notin \{3, 12, 21, 106, 203\}, \sum_j x_j = 1\},$$

- In some sense, k-sparse model should a $k+1$-face of a combinatorial structure
- *If fewer than 'full' number of faces* of a projected polytope *can't determine* certain models from data.

Underdetermined Equations w/ nonnegative coefficients

Solution method:

$$(NP) \qquad x_0 = \operatorname{argmin} \|x\|_0 \text{ s.t. } y = Ax, \quad x \geq 0$$

$$(LP) \qquad x_1 = \operatorname{argmin} \|x\|_1 \text{ s.t. } y = Ax, \quad x \geq 0$$

Lemma Suppose that x_0 is a fixed k-sparse vector, and A is Gaussian iid $N(0, 1/n)$. Then

$$P\{x_1 = x_0\} = \frac{f_{k-1}(AT^{N-1})}{f_{k-1}(T^{N-1})}$$

Theorem For large N, n, $n/N \to \delta$ and $k/n \to \rho$

$$\frac{f_{k-1}(AT^{N-1})}{f_{k-1}(T^{N-1})} \geq 1 - o(1)$$

provided $\rho < \rho_W^+(\delta)$. DLD and Tanner 2005.

Underdetermined Equations

Solution method:

$$(P_0) \qquad x_0 = \operatorname{argmin} \|x\|_0 \text{ s.t. } y = Ax$$

$$(P_1) \qquad x_1 = \operatorname{argmin} \|x\|_1 \text{ s.t. } y = Ax$$

Lemma Suppose that x_0 is a fixed k-sparse vector. Consider random problem instance $y = Ax_0$, A Gaussian $N(0, 1/n)$. Then

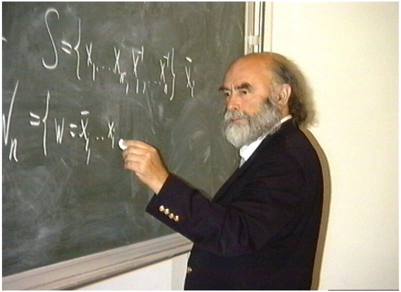
$$P\{x_1 = x_0\} = \frac{f_{k-1}(AC^N)}{f_{k-1}(C^N)}.$$

Theorem For large N, n, $n/N \to \delta$ and $k/n \to \rho$

$$\frac{f_{k-1}(AC^N)}{f_{k-1}(C^N)} \geq 1 - o(1)$$

provided $\rho < \rho_W^{\pm}(\delta)$. DLD (2004,2005)

A. Vershik & R. Schneider

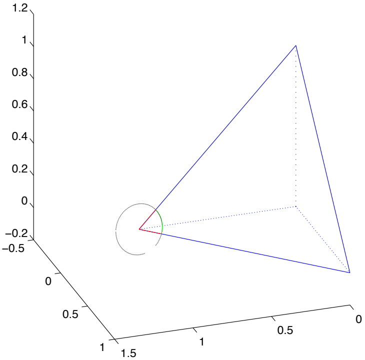


Expected Face Count

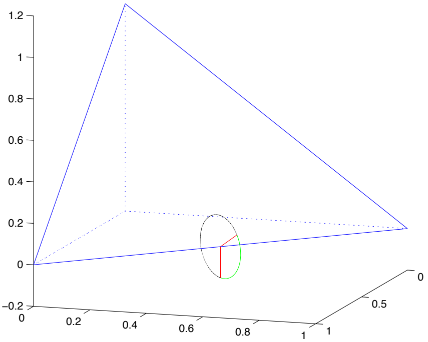
$$Ef_k(R) = f_k(Q) - 2\sum_{s\geq 0}\sum_{F\in\mathcal{F}_k(Q)}\sum_{G\in\mathcal{F}_{n+1+2s}(Q)} \beta(F,G)\gamma(G,Q);$$

- $R = AQ$ with A uniform random projection
- Affentranger and Schneider (1992), Vershik and Sproyshev (1992) McMullen (1972), Grünbaum (1968) H. Ruben (1960)
- $\beta(F,G)$ Internal Angles (explain)
- $\gamma(G,Q)$ External Angles (explain)

Example of Internal Angle



Example of External Angles



Angles and Gaussian Integrals

$$\gamma(F^{\ell}, T^{n-1}) = \sqrt{\frac{\ell+1}{\pi}} \int_0^{\infty} e^{-(\ell+1)x^2} \left(\frac{2}{\sqrt{\pi}} \int_0^{x} e^{-y^2} dy \right)^{n-\ell-1} dx.$$

$\beta(F, G)$ is proportional to:

$$J(m, \theta) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \left(\int_0^{\infty} e^{-\theta v^2 + 2iv\lambda} dv \right)^m e^{-\lambda^2} d\lambda.$$

Asymptotics as $\delta \to 0$

Theorem DLD & Tanner JAMS 2008

$$\rho_W^+(\delta) \sim \rho_W^\pm(\delta) \sim \frac{1}{2\log(1/\delta)}, \qquad \delta \to 0.$$

$$\rho_S^+(\delta) \sim \frac{1}{2e|\log(2\sqrt{\pi}\delta)|}, \qquad \delta \to 0.$$

$$\rho_S^\pm(\delta) \sim \frac{1}{2e|\log(\sqrt{\pi}\delta)|}, \qquad \delta \to 0.$$

Proof: Careful analysis of internal, external angles in asymptotic setting where $d, k, \ell \ll n$.

Practical Implication of $\delta \longrightarrow 0$: Undersampling Theorem

Rule of Thumb.

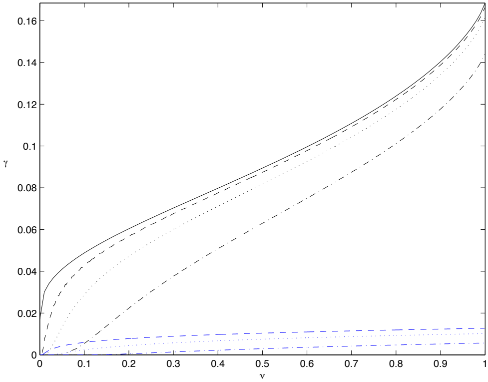
- Suppose k, N both large.
- k/N is small and $x_0 \in \mathbf{R^n}$ is k-sparse.
- How many random projections of x_0 do we need so that (P_1) perfectly reconstructs?
- Answer:

$$n \approx 2k \log(N/n).$$

- Proof: $\rho^{\pm}_{W}(n/N) \approx \frac{1}{2\log(n/N)}$ for n/N small.
- Implication: Compressed Sensing. DLD IEEE IT 2006

Recent Developments: Finite n bounds (a)

DLD & Tanner (2008) ; $n = \infty$, 5000, 1000,200



Recent Developments: Finite n bounds (b)

DLD & Tanner (2008) ;

	Positivity-Constrained				Unconstrained		
N	n	k	ϵ	θ	n	k	
10^4	3,529	1,253	10^{-3}	$1/5$	4,299	1,208	10
10^6	30,510	4,472	10^{-3}	$1/10$	31,323	3860	10
10^6	35,766	5,487	10^{-10}	$1/10$	36,819	4,722	10^-
10^9	1,355,580	113,004	10^{-10}	$1/50$	1,365,079	102,646	10^-

Table: For the specified (k, n, N) and ϵ, the probability of succesfully recovering a k-sparse vector $x_0 \in \mathbf{R^N}$ from n samples exceeds $1 - \epsilon$. θ is a parameter of our method measuring proximity to the asymptotic thresholds

Recent Developments: Finite n bounds (c)

DLD & Tanner (2008)

	Here		RV	
N	n	k	n	k
10^3	500	40	652	1
10^4	1,000	45	1,089	3
10^5	1,000	27	4,377	27
10^7	100,000	2,993	100,090	895

Table: Comparison of our existence results with implications of Rudelson and Vershynin. Note: RV seemingly much stronger than RIP arguments of Candes-Tao

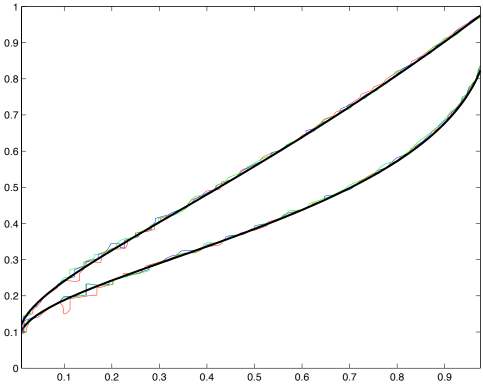
Recent Developments: Universality

- Same experiment, different matrix ensembles
  - Bernoulli, Rademacher
  - Partial Fourier, Hadamard: sample n rows from $N \times N$ ortho matrix
  - Gaussian: $A_{i,j} \sim N(0, 1)$
  - Uniform Random Projection
  - Expander Graphs

Empirically: same behavior. Tanner and DLD (2009, Submitted Phil. Trans. Roy. Soc.)

- 16,000 situations studied
- 2.8 Million Linear programs solved
- 6.8 CPU years
- Data analysis: 33 pages

Results for 9 Matrix Ensembles

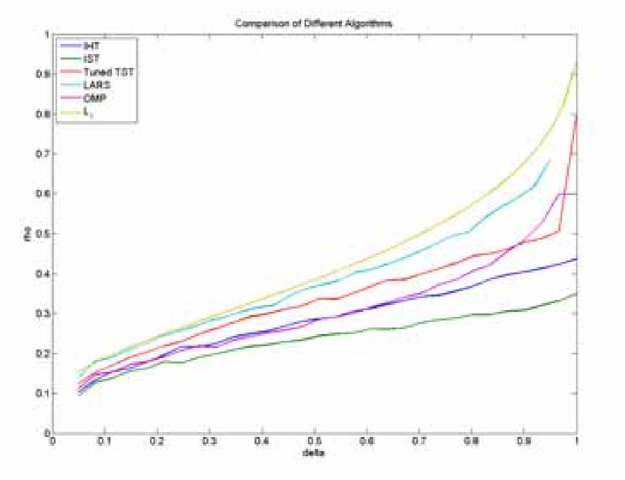


$N = 1600$. Lower Curve, Coeff. $\pm$; Upper Curve, Coeff. $+$.

Other Algorithms

- Basic Schemas
  - Iterative Hard/Soft Thresholding
  - Iterative Multiple Regression
- Tuning
  - Threshold Levels
  - Relaxation Parameters
  - Pruning Parameters
- Empirical Study:
  Maleki and DLD (2009, Submitted IEEE SP)
  - 100 Million Reconstructions
  - 3.8 CPU years
  - Vary Matrix & Coefficient ensembles

Maximin Tuned Algorithms Compared with ℓ_1



Is this Fundamental?

- We've been studying *The Hydrogen Atom*.
  - no noise
  - exact sparsity
  - ℓ_1 only
- 'Natural Boundary' for more general settings
- Relax Exact Sparsity: Xu and Hassibi (2008)
- Add noise, Minimax Bayes Risk (w/Johnstone)
  - Add noise, let $\delta = n/N \rightarrow zero$: minimax MSE explodes above $2k \log(N/n)(1 + o(1))$.
- Other properties, story can differ, see Wainwright and co-authors.
- Other polytopes & convex bodies: eg Hypercube, etc. DLD & Tanner; Semidefinite Cone Recht, Xu, Hassibi

Take Away Messages

- Draw Phase diagrams: n/N on one axis k/n on other.
- Phase Transitions:
  - Observable empirically
  - Rigorous foundation in combinatorial geometry
  - Universal across many matrix ensembles
  - Exist for many algorithms, properties
- Phase Transition at $n \asymp 2k \cdot \log(N/n)$ is useful, fundamental
- zero-noise, exact sparsity PT continued analytically to positive-noise, approximate sparsity results

Papers Mentioned

- DLD & JW Tanner (2005) Sparse Nonnegative Solution of Underdetermined Linear Systems. *PNAS* **102** 9446-9451.
- DLD & JW Tanner (2005) Neighborliness of Randomly Projected Simplices in High Dimensions. *PNAS* **102** 9452-9457.
- DLD (2005) High-Dimensional Centrosymmetric Polytopes with Neighborliness Proportional to Dimension. *Discrete and Computational Geometry.* Online, Dec. 2005.
- DLD (2006) Compressed Sensing. IEEE IT.
- DLD & JW Tanner (2006) Counting Faces of Randomly Projected Polytopes when the Projection Radically Lowers Dimension. Journ. Amer Math Soc, Online July 2008
- DLD & JW Tanner (2008) Exponential Bounds Implying Construction of Compressed Sensing Matrices, Error-Correcting Codes and Neighborly Polytopes by Random Sampling submitted
- DLD & Y Tsaig (2008) Fast solution of ℓ_1 minimization problems when the solution may be sparse. IEEE IT, Nov. 2008.
- DLD & JW Tanner (2009) Observed Universality of Phase Transitions associated to Compressed Sensing, Error-Correcting Codes, and Neighborly Polytopes. submitted.
- A. Maleki and DLD (2009) Optimally-tuned, freely-available Iterative Reconstruction Algorithms for Compressed Sensing. submitted.
- DLD & JW Tanner (2009) Precise Undersampling Theorems. submitted.