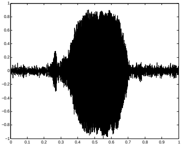
Random matrix theory and stochastic geometry in CS

Jared Tanner

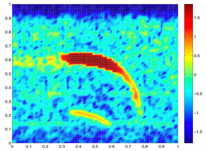
Probability and geometry in high dimensions
University of Edinburgh
Joint with Bah, Blanchard, Cartis, and Donoho

Compressed Sensing - Encoder

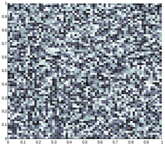
- *Data acquisition at the information rate*
  - When it is "costly" to acquire information use CS
  - Transform workload from sensor to computing resources
  - Reduced sampling possible by exploiting simplicity
- Linear Encoder: Discrete signal of length N, x
  - Transform matrix under which class of signals are sparse, Φ
  - "Random" matrix to mix transform coefficients, A
  - Measurements through $A\Phi$, $n \times N$ with $n \ll N$, $b := A\Phi x$



x



Φx



"row" of A

Compressed Sensing - Decoder

- *Data acquisition at the information rate*
  - When it is "costly" to acquire information use CS
  - Transform workload from sensor to computing resources
  - Reduced sampling possible by exploiting simplicity
- Linear Encoder: Discrete signal of length N, x
  - Transform matrix under which class of signals are sparse, Φ
  - "Random" matrix to mix transform coefficients, A
  - Measurements through $A\Phi$, $n \times N$ with $n \ll N$, $b := A\Phi x$
- Decoder: Reconstruct an approximation of x from (b, A)
  - Thresholding: take large coefficients of $A^* b$
  - Greedy Algorithms: OMP, CoSaMP, SP, IHT, StOMP, ...
  - Regularization: $\min_y \|\Phi y\|_1$ subject to $\|A\Phi y - b\|_2 \leq \eta$

Sparse Approximation Phase Transitions

- Problem characterized by three numbers: $k \leq n \leq N$
  - N, Signal Length, "Nyquist" sampling rate
  - n, number of inner product measurements
  - k, signal complexity, sparsity

Sparse Approximation Phase Transitions

- Problem characterized by three numbers: $k \leq n \leq N$
  - N, Signal Length, "Nyquist" sampling rate
  - n, number of inner product measurements
  - k, signal complexity, sparsity
- For what (k, n, N) does an encoder/decoder pair recover a suitable approximation of x from (b, A)?
  - $n \sim k^2$ sufficient for many encoder/decoder pairs
  - $n = k$ is the optimal oracle rate
  - $n \sim k$ possible using computationally efficient algorithms

Sparse Approximation Phase Transitions

- Problem characterized by three numbers: $k \leq n \leq N$
  - N, Signal Length, "Nyquist" sampling rate
  - n, number of inner product measurements
  - k, signal complexity, sparsity
- For what (k, n, N) does an encoder/decoder pair recover a suitable approximation of x from (b, A)?
  - $n \sim k^2$ sufficient for many encoder/decoder pairs
  - $n = k$ is the optimal oracle rate
  - $n \sim k$ possible using computationally efficient algorithms
- Mixed *under/over-sampling* rates compared to naive/optimal

$$\text{Undersampling: } \delta := \frac{n}{N}, \qquad \text{Oversampling: } \rho := \frac{k}{n}$$

Methods of Analysis: conditions on *encoder*

- Generic measures of used to imply algorithm success:
  - Coherence: maximum correlation of columns, $\max_{i \neq j} |a_i^* a_j|$
  - Restricted Isometry Property (RIP): sparse near isometry

$$(1 - R_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + R_k)\|x\|_2^2 \quad \text{for } x \;\; k\text{-sparse}$$

ℓ^1-regularization "works" if $R_{2k} < 0.45$ (Foucart & Lai)

Methods of Analysis: conditions on *encoder*

- Generic measures of used to imply algorithm success:
  - Coherence: maximum correlation of columns, $\max_{i\neq j}|a_i^* a_j|$
  - Restricted Isometry Property (RIP): sparse near isometry

$$(1-R_k)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+R_k)\|x\|_2^2 \quad \text{for } x \ \ k\text{-sparse}$$

ℓ^1-regularization "works" if $R_{2k} < 0.45$ (Foucart & Lai)

- Algorithm specific:
  - False Alarm/Discovery Rate: Stagewise OMP (StOMP)
  - Convex Polytopes (face counting): ℓ^1-regularization

Methods of Analysis: conditions on *encoder*

- Generic measures of used to imply algorithm success:
  - Coherence: maximum correlation of columns, $\max_{i \neq j} |a_i^* a_j|$
  - Restricted Isometry Property (RIP): sparse near isometry

$$(1 - R_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + R_k)\|x\|_2^2 \quad \text{for } x \;\; k\text{-sparse}$$

  ℓ^1-regularization "works" if $R_{2k} < 0.45$ (Foucart & Lai)

- Algorithm specific:
  - False Alarm/Discovery Rate: Stagewise OMP (StOMP)
  - Convex Polytopes (face counting): ℓ^1-regularization
- Measures of success:
  - Success for *all* k-sparse signals (RIP, polytopes)
  - Success for *most* signals (coherence, FAR, polytopes)

Restricted Isometry Constants (RIC)

- Restricted Isometry Constants (RIC): for all k-sparse x

$$(1 - L(k, n, N; A))\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + U(k, n, N; A))\|x\|_2^2$$

- Most sparsity algorithms have optimal recovery rate if RICs remain bounded as $k/n \to \rho$, $n/N \to \delta$, with $\rho, \delta \in (0, 1)$.
- What do we know about bounds on RICs?

Restricted Isometry Constants (RIC)

- Restricted Isometry Constants (RIC): for all k-sparse x

$$(1 - L(k,n,N;A))\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + U(k,n,N;A))\|x\|_2^2$$

- Most sparsity algorithms have optimal recovery rate if RICs remain bounded as $k/n \to \rho$, $n/N \to \delta$, with $\rho, \delta \in (0,1)$.
- What do we know about bounds on RICs?
- No known large deterministic rect. matrices with bounded RIC
- Ensembles with concentration of measure have bounded RIC

$$P(|\|Ax\|_2^2 - \|x\|_2^2| \geq \epsilon\|x\|_2^2) \leq e^{-n \cdot c(\epsilon)} \quad c(\epsilon) > 0.$$

Gaussian, uniform $\{-1, 1\}$, most any with i.i.d. mean zero

Restricted Isometry Constants (RIC)

- Restricted Isometry Constants (RIC): for all k-sparse x

$$(1 - L(k,n,N;A))\|x\|_2^2 \le \|Ax\|_2^2 \le (1 + U(k,n,N;A))\|x\|_2^2$$

- Most sparsity algorithms have optimal recovery rate if RICs remain bounded as $k/n \to \rho$, $n/N \to \delta$, with $\rho, \delta \in (0,1)$.
- What do we know about bounds on RICs?
- No known large deterministic rect. matrices with bounded RIC
- Ensembles with concentration of measure have bounded RIC

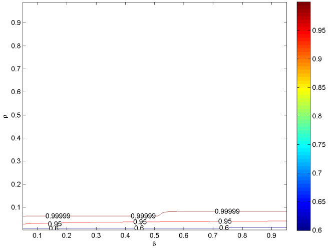
$$P(|\|Ax\|_2^2 - \|x\|_2^2| \ge \epsilon \|x\|_2^2) \le e^{-n \cdot c(\epsilon)} \quad c(\epsilon) > 0.$$

  Gaussian, uniform $\{-1, 1\}$, most any with i.i.d. mean zero
- How large are these RICs? When do we have guarantees for sparsity recovery? $\max(U(k,n,N;A), L(k,n,N;A)) \le \sqrt{2} - 1$

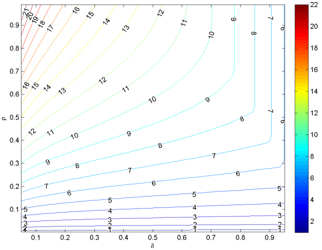
Random Matrix Theory and the RIC

- RIC bounds for Gaussian $\mathcal{N}(0, n^{-1})$ [Candés and Tao 05]

$$(1 - L(\delta,\rho))\|x\|_2^2 \le \|Ax\|_2^2 \le (1 + U(\delta,\rho))\|x\|_2^2$$



$L(\delta,\rho)$



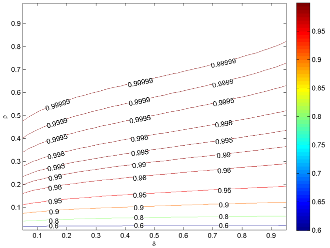
$U(\delta,\rho)$

- Always stated as "$\delta_k := \max(L(k,n,N;A), U(k,n,N;A))$"
- Bound: concentration of measure + $\binom{N}{k}$ union bound

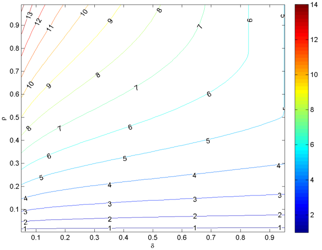
Random Matrix Theory and the RIC

- RIC bounds for Gaussian $\mathcal{N}(0, n^{-1})$ [Bl-Ca-Ta 09]

$$(1 - L(\delta,\rho))\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + U(\delta,\rho))\|x\|_2^2$$



$L(\delta,\rho)$



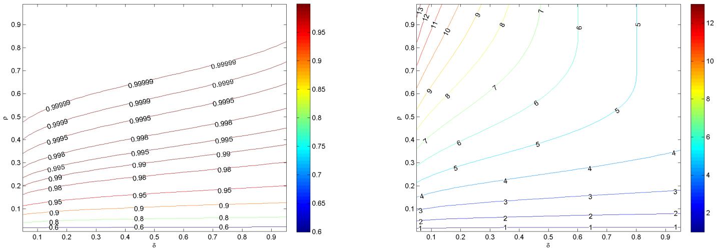
$U(\delta,\rho)$

- First asymmetric bounds, dramatic improvement for $L(\delta,\rho)$
- Bound: Large deviation of Wishart PDFs + $\binom{N}{k}$ union bound

Random Matrix Theory and the RIC

- RIC bounds for Gaussian $\mathcal{N}(0, n^{-1})$ [Bah-Ta 10]

$$(1 - L(\delta,\rho))\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + U(\delta,\rho))\|x\|_2^2$$



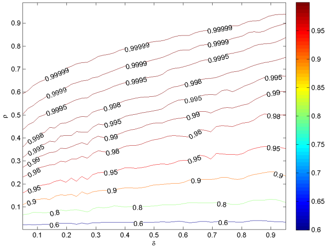
$L(\delta,\rho)$ $\qquad$ $U(\delta,\rho)$

- Exploit eigenvalue "smoothness" for overlapping submatrices
- No more than 1.57 times empirically observations values

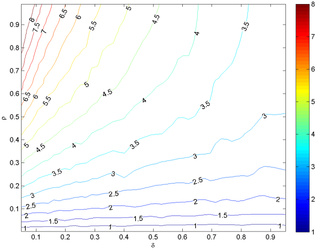
Random Matrix Theory and the RIC

- Observed RIC for Gaussian $\mathcal{N}(0, n^{-1})$ [Bah-Ta 09]

$$(1 - L(k,n,N))\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + U(k,n,N))\|x\|_2^2$$



$L(k, n, N)$



$U(k, n, N)$

- Observed lower bounds for $n = 400$ and various (k, N)
- What do these RICs tell us for sparsity algorithms?

Algorithms for Sparse Approximation

Input: A, b, and possibly tuning parameters

- ℓ^1-regularization:

$$\min_x \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \tau$$

- Simple Iterated Thresholding:

$$x^{t+1} = H_k(x^t + \kappa A^T(b - Ax^t))$$

- Two-Stage Thresholding (Subspace Pursuit, CoSaMP):

$$v^{t+1} = x^{t+1} = H_{\alpha k}(x^t + \kappa A^T(b - Ax^t))$$

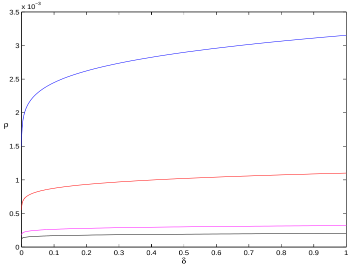
$$I_t = supp(v^t) \cup supp(x^t) \quad \text{Join supp. sets}$$

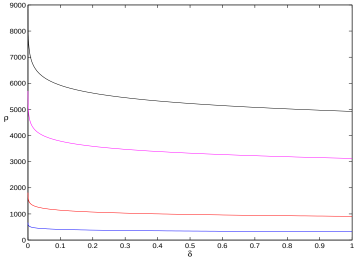
$$w_{I_t} = (A_{I_t}^T A_{I_t})^{-1} A_{I_t}^T b \quad \text{Least squares fit}$$

$$x^{t+1} = H_{\beta k}(w^t) \quad \text{Second threshold}$$

When does RIP guarantee they work?

Best known bounds implied by RIP



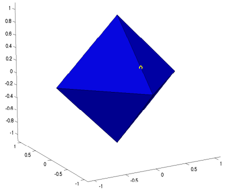


- Lower bounds on the Strong exact recovery phase transition for Gaussian random matrices for the algorithms ℓ^1-regularization, IHT, SP, and CoSaMP (black).
  - Unfortunately recovery thresholds are impractically low.

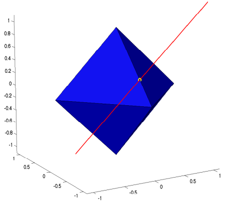
  $n > 317k$, $n > 907k$, $n > 3124k$, $n > 4925k$
- Larger phase transitions appear only possible by using algorithm specific techniques of analysis.

Geometry of ℓ^1-regularization, $\mathbb{R}^N$

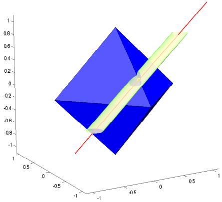
- Sparsity: $x \in \mathbb{R}^N$ with $k < n$ nonzeros on $k-1$ face of ℓ^1 ball.
- Null space of A intersects C^N at only x, or pierces C^N



ℓ^1 ball $\in \mathbb{R}^N$



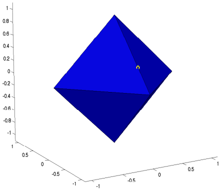
$x + \mathcal{N}(\mathcal{A})$



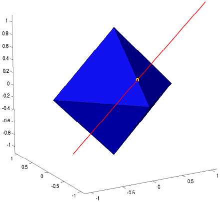
$\|A(x-y)\| \leq \eta$

Geometry of ℓ^1-regularization, $\mathbb{R}^N$

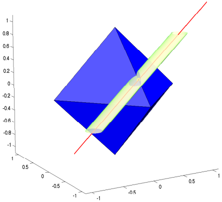
- Sparsity: $x \in \mathbb{R}^N$ with $k < n$ nonzeros on $k-1$ face of ℓ^1 ball.
- Null space of A intersects C^N at only x, or pierces C^N



ℓ^1 ball $\in \mathbb{R}^N$



$x + \mathcal{N}(\mathcal{A})$

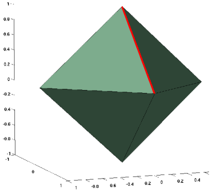


$\|A(x-y)\| \leq \eta$

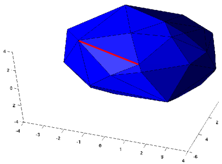
- If $\{x + \mathcal{N}(\mathrm{A})\} \bigcap C^N = x$, ℓ^1 minimization recovers x
- Faces pierced by $x + \mathcal{N}(\mathcal{A})$ do not recover k sparse x

Geometry of ℓ^1-regularization, $\mathbb{R}^n$

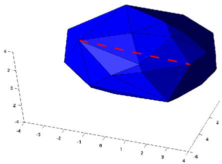
- Sparsity: $x \in \mathbb{R}^N$ with $k < n$ nonzeros on $k-1$ face of ℓ^1 ball.
- Matrix A projects face of ℓ^1 ball either onto or into $conv(\pm A)$.



ℓ^1 ball $\in \mathbb{R}^N$



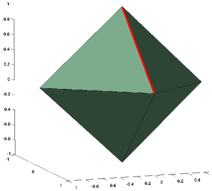
edge onto $conv(\pm A)$



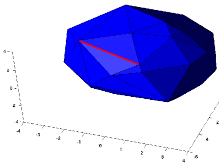
edge into $conv(\pm A)$

Geometry of ℓ^1-regularization, $\mathbb{R}^n$

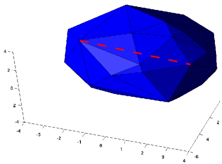
- Sparsity: $x \in \mathbb{R}^N$ with $k < n$ nonzeros on $k-1$ face of ℓ^1 ball.
- Matrix A projects face of ℓ^1 ball either onto or into $conv(\pm A)$.



ℓ^1 ball $\in \mathbb{R}^N$



edge onto $conv(\pm A)$

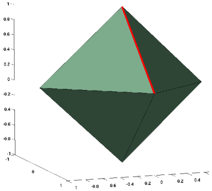


edge into $conv(\pm A)$

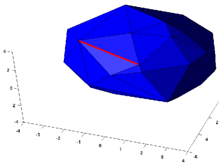
- Survived faces are sparsity patterns in x where $\ell^1 \to \ell^0$
- Faces which fall inside $conv(\pm A)$ are not solutions to ℓ^1

Geometry of ℓ^1-regularization, $\mathbb{R}^n$

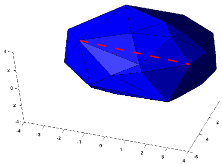
- Sparsity: $x \in \mathbb{R}^N$ with $k < n$ nonzeros on $k-1$ face of ℓ^1 ball.
- Matrix A projects face of ℓ^1 ball either onto or into $conv(\pm A)$.



ℓ^1 ball $\in \mathbb{R}^N$



edge onto $conv(\pm A)$



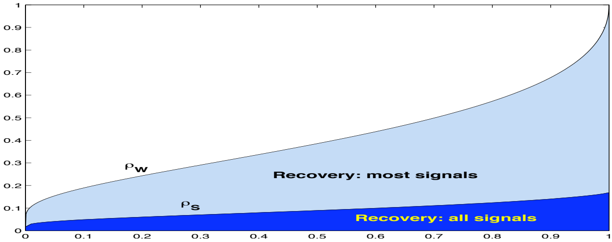
edge into $conv(\pm A)$

- Survived faces are sparsity patterns in x where $\ell^1 \to \ell^0$
- Faces which fall inside $conv(\pm A)$ are not solutions to ℓ^1
- Neighborliness of random polytopes [Affentranger & Schneider]
- Exact recoverability of k sparse signals by "counting faces"

Phase Transition: ℓ^1 ball, C^N

- With overwhelming probability on measurements $A_{n,N}$: for any $\epsilon > 0$, as $(k, n, N) \to \infty$
  - *All* k-sparse signals if $k/n \le \rho_S(n/N, C)(1-\epsilon)$
  - *Most* k-sparse signals if $k/n \le \rho_W(n/N, C)(1-\epsilon)$
  - *Failure typical* if $k/n \ge \rho_W(n/N, C)(1+\epsilon)$

$\frac{k}{n}$

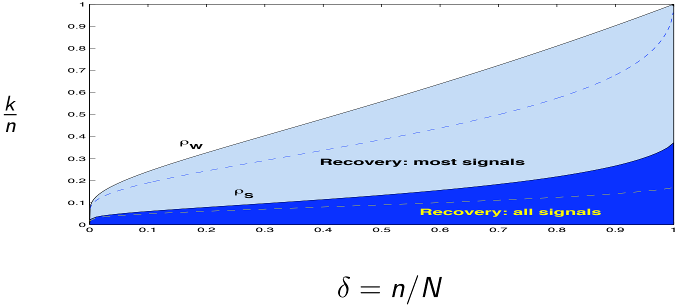


$\delta = n/N$

- Asymptotic behavior $\delta \to 0$: $\rho(n/N) \sim [2(e)\log(N/n)]^{-1}$

Phase Transition: Simplex, T^{N-1}, $x \geq 0$

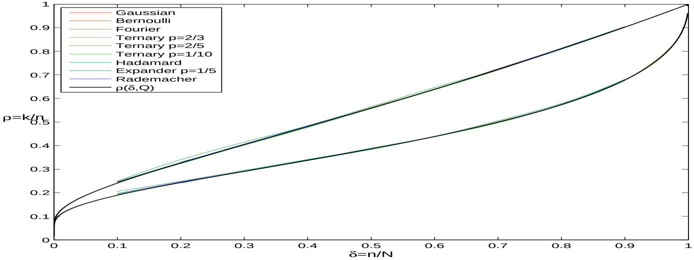
- With overwhelming probability on measurements $A_{n,N}$:
  for any $\epsilon > 0$, $x \geq 0$, as $(k, n, N) \to \infty$
  - *All* k-sparse signals if $k/n \leq \rho_S(n/N, T)(1-\epsilon)$
  - *Most* k-sparse signals if $k/n \leq \rho_W(n/N, T)(1-\epsilon)$
  - *Failure typical* if $k/n \geq \rho_W(n/N, T)(1+\epsilon)$



- Asymptotic behavior $\delta \to 0$: $\rho(n/N) \sim [2(e)\log(N/n)]^{-1}$

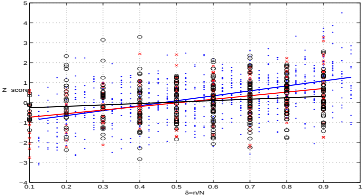
Weak Phase Transitions: Visual agreement

- Black: Weak phase transition: $x \geq 0$ (top), x signed (bot.)
- Overlaid empirical evidence of 50% success rate:

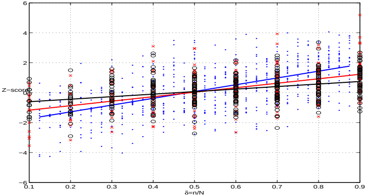


- Gaussian, Bernoulli, Fourier, Hadamard, Rademacher
- Ternary (p): $P(0) = 1 - p$ and $P(\pm 1) = p/2$
- Expander (p): $\lceil p \cdot n \rceil$ ones per column, otherwise zeros
- Rigorous statistical comparison shows $N^{-1/2}$ convergence

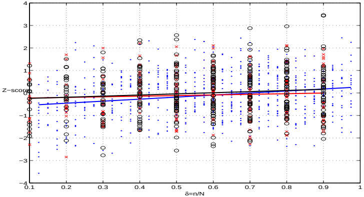
Bulk Z-scores



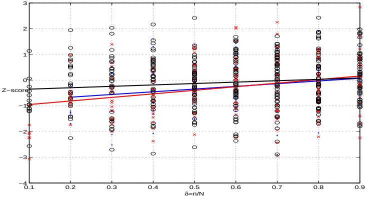
(a) Bernoulli



(b) Fourier



(c) Ternary (1/3)

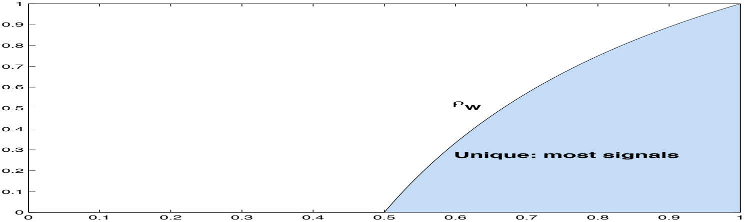


(d) Rademacher

- $N = 200$, $N = 400$ and $N = 1600$
- Linear trend with $\delta = n/N$, decays at rate $N^{-1/2}$

Phase Transition: Hypercube, H^N

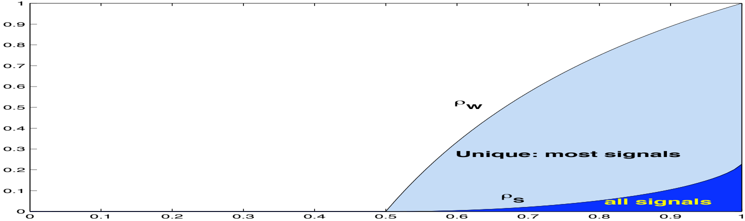
- Let $0 \leq x \leq 1$ have k entries $\neq 0, 1$ and form $b = Ax$.
- Are there other $y \in H^N[0,1]$ such that $Ay = b$, $y \neq x$?
- As $n, N \to \infty$, Typically No provided $k/n < \rho_W(\delta; H)$



- Unlike T and C: no strong phase transition
- Universal: A need only be in general position
- Simplicity beyond sparsity: Hypercube k-faces correspond to vectors with only k entries away from the bounds (not 0 or 1).

Phase Transition: Orthant, $\mathbb{R}^N_+$

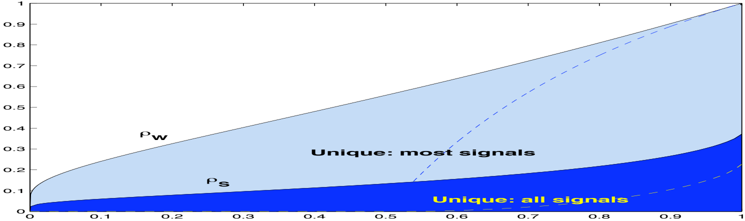
- Let $x \geq 0$ be k-sparse and form $b = Ax$.
- Are there other $y \in \mathbb{R}^N_+$ such that $Ay = b$, $y \geq 0$, $y \neq x$?
- As $n, N \to \infty$, Typically No provided $k/n < \rho_W(\delta; \mathbb{R}_+)$



- Universal: A columns centrally symmetric and exchangeable
  Not universal to all A in general position–design possible.
- For $k/n < \rho_W(\delta, \mathbb{R}_+) := [2 - 1/\delta]_+$ and $x \geq 0$,
  any "feasible" method will work, e.g. WCP (Cartis & Gould)

Phase Transition: Orthant, $\mathbb{R}^N_+$, matrix design

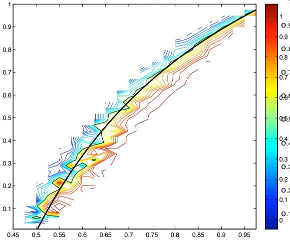
- Let $x \geq 0$ be k-sparse and form $b = Ax$.
- Are there other $y \in \mathbb{R}^N_+$ such that $Ay = b$, $y \geq 0$, $y \neq x$?
- As $n, N \to \infty$, Typically No provided $k/n < \rho_W(\delta; \mathbb{R}_+)$



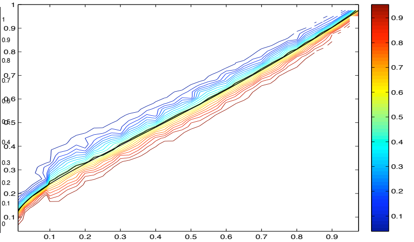
- Gaussian and measuring the mean (row of ones):
  $\rho_W(n/N; \mathbb{R}_+) \to \rho_W(n/N; T)$
- Simple modification of A makes profound difference
  Unique even for $n/N \to 0$ with $n > 2(e)k \log(N/n)$

Orthant matrix design, it's really true

- Let $x \geq 0$ be k-sparse and form $b = Ax$.
- Not ℓ^1, but: $\max_y \|x - y\|$ subject to $Ay = Ax$ and $y \geq 0$
- Good empirical agreement for $N = 200$.



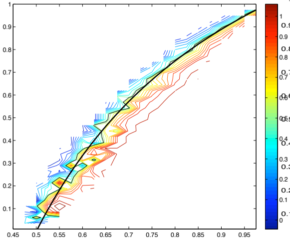
Rademacher



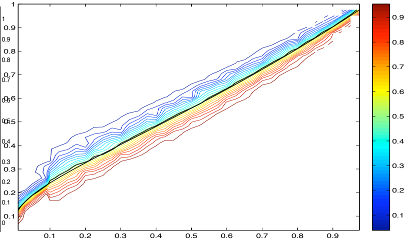
Rademacher and row of ones

Orthant matrix design, it's really true

- Let $x \geq 0$ be k-sparse and form $b = Ax$.
- Not ℓ^1, but: $\max_y \|x - y\|$ subject to $Ay = Ax$ and $y \geq 0$
- Good empirical agreement for $N = 200$.



Rademacher



Rademacher and row of ones

SUMMARY	Simplex	ℓ^1 ball	Hypercube	Orthant
Matrix class	Gaussian	Gaussian	gen. pos.	sym. exch.
Design	Vandermonde	unknown	not possible	row ones

References

- Donoho (2005) High-Dimensional Centrosymmetric Polytopes with Neighborliness Proportional to Dimension. *DCG*.
- Donoho and Tanner
  - Counting faces of randomly-projected polytopes when projection radically lowers dimension (2006), *J. AMS*.
  - Counting the faces of randomly projected hypercubes and orthants, with applications (2008), *DCG*.
  - Precise Undersampling Theorems (2009), *Proc. IEEE*.
  - Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing (2009) *Phil. Trans. Roy. Soc. A*.
- Blanchard, Cartis, Tanner, Thompson (2009) Compressed Sensing: How sharp is the RIP?
- Bah, Tanner (2010) Improved RIC bounds for Gaussian Matrices.

Thanks for your time