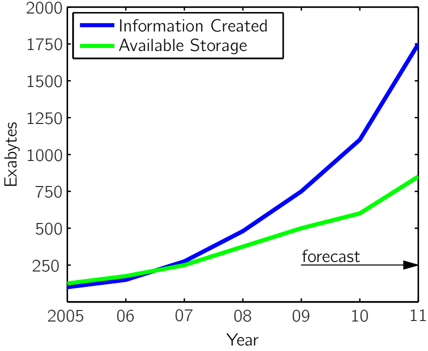
# Applications of concentration of measure in signal processing

Tyrone Vincent, Michael Wakin and Luis Tenorio

Colorado School of Mines

April 30, 2010

# Signal Processing in the Age of the Data Flood

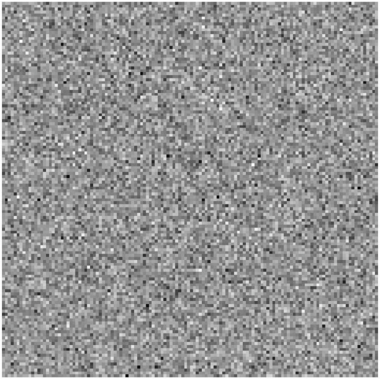


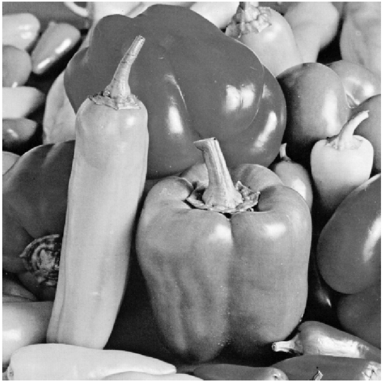

Reference: Economist magazine, Feb 25, 2010.

- Exabyte $= 2^{60}$ bits.
- We have passed the point where all data created can be stored
- LHC generates $40$ Tb every second.
- Other bottlenecks
  - acquisition
  - transmission
  - analysis

# Not all length-$N$ signals are created equal

- What is the class of "typical images"?



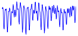


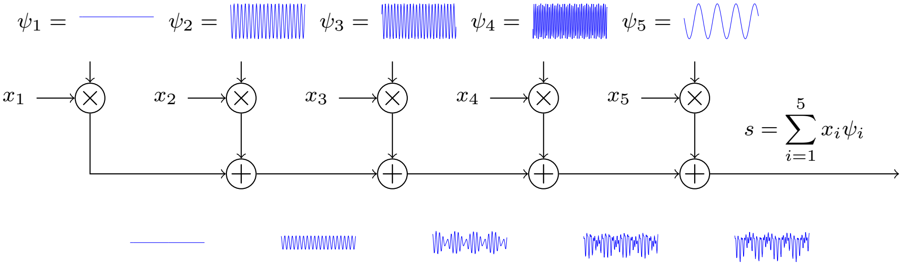
- "Typical" signals contain degrees of freedom $S$ less than $N$

# Dimensionality Reduction

- Can we reduce the burden from $N$ to $S$ early and often in the data processing pipeline?

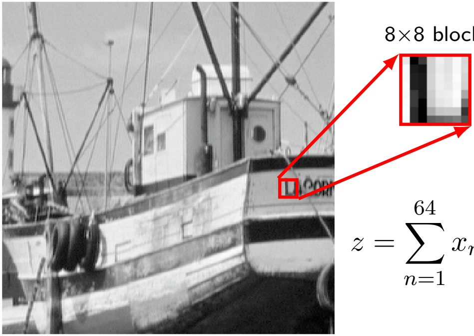
# Signal Representation: Signal Basis

- A signal basis can be used to define the class of signals of interest
- Example: represent a signal $z =$  as sum of scaled sinusoids



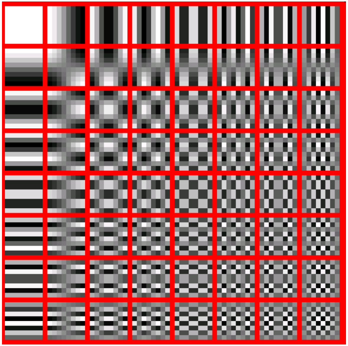
# Lossy Compression: JPEG

Discrete Cosine Signal Basis $\psi_k$



8×8 block

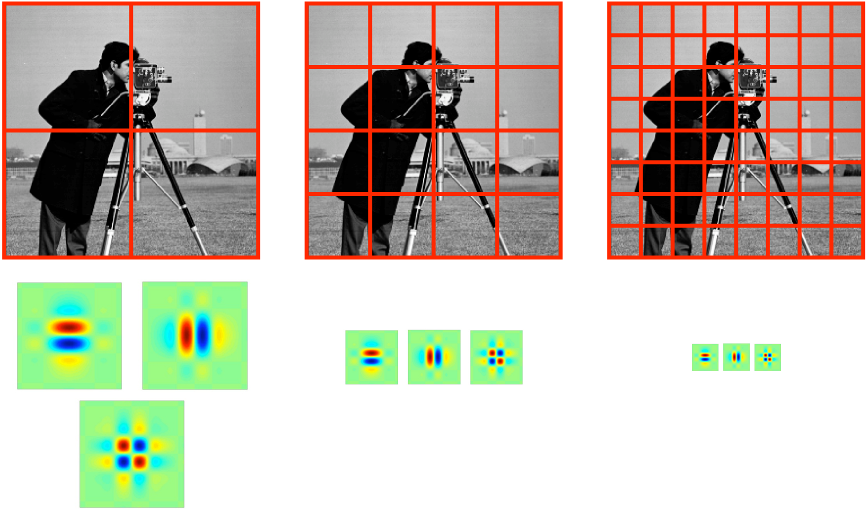
$$z = \sum_{n=1}^{64} x_n \psi_n$$



credit: J. Romberg

- Approximation with quantized coefs: $\hat{z} = \sum_{n=1}^{64} \hat{x}_n \psi_n$;

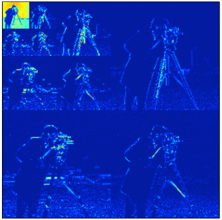
# Multi-Scale Basis: Wavelets

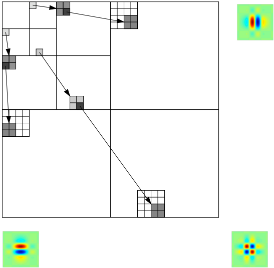


credit: J. Romberg

# Wavelet coefficient representation

credit: J. Romberg



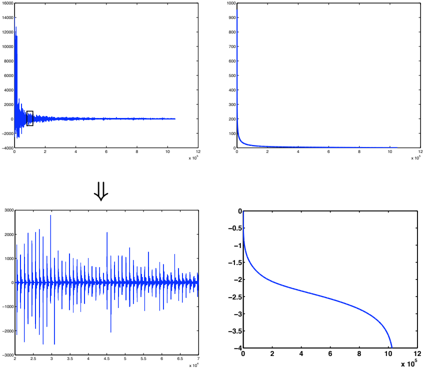


- A few large coefficients, but many small coefficients.

# How many coefficients are important?

1 megapixel image

wavelet coeffs (sorted)

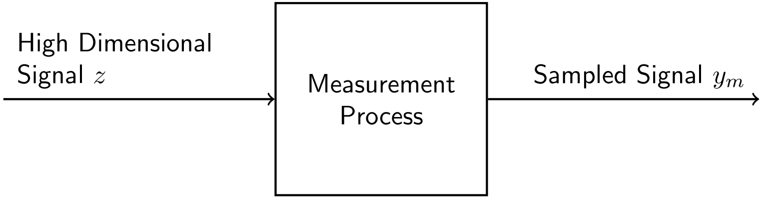


zoom in ($\log_{10}$ sorted)

# Conclusion

- Many classes of signals have a *sparse* representation in an appropriate basis

# Now let's bring in the measurement

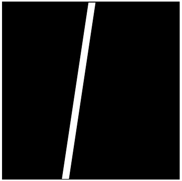


# Measurement Models

- Many measurement modalities are *linear*
- Inner product representation:

$$y_m = \langle z, \phi_m \rangle = \text{sum of point-wise product}$$

- Tomography

$y_m = \langle$  ,  $\rangle$

credit: J. Romberg

# Band-limited Signal Recovery:

## Signal and Measurement Model

- The signal/measurement model for the (1-D) Nyquist theorem uses
  - Signal model basis:

    $$\psi_n = e^{j\omega_0 n t}$$

  - Measurement: Sampling, $M$ samples per period $T_s = T_0/M$.

    $$\phi_m = \delta(t - T_s m)$$

    Sampling frequency is $f_s = \frac{1}{T_s} = \frac{M}{T_0} = M\omega_0$.
  - *a priori* information: Band-limited signal, i.e. coefficients zero for $|n| \geq N_b$. Bandwidth: $\omega_b = N_b \omega_0$

# Band-limited Signal Recovery:
## Set of Linear Equations

- Using this model,

$$y_m = \left\langle \sum_{n=-N_b}^{N_b} x_n \psi_n, \phi_m \right\rangle$$

$$y_m = \begin{bmatrix} & a_m' & \end{bmatrix} \begin{bmatrix} x_{-N_b} \\ x_{1-N_b} \\ \vdots \\ x_{N_b-1} \\ x_{N_b} \end{bmatrix}$$

where $a_k = \begin{bmatrix} \langle \psi_{-N_b}, \phi_m \rangle & \cdots & \langle \psi_{N_b}, \phi_m \rangle \end{bmatrix}$

# Band-limited Signal Recovery:

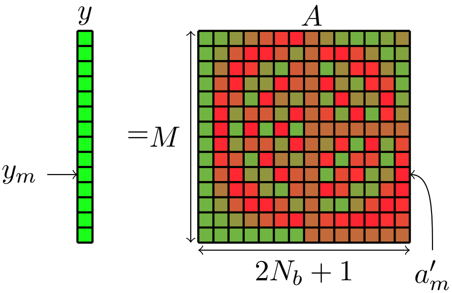
## Row Independence

$$
\begin{aligned}
a_k &= \begin{bmatrix} \langle \psi_{-N_b}, \phi_m \rangle & \cdots & \langle \psi_{N_b}, \phi_m \rangle \end{bmatrix} \\
&= \begin{bmatrix} e^{j\omega_0 T_s m(-N_b)} & e^{j\omega_0 T_s m(-N_b+1)} & \dots & e^{j\omega_0 T_s m(N_b)} \end{bmatrix}
\end{aligned}
$$

- $a_k$ looks like $e^{j\hat{\omega}n}$ with $\hat{\omega} = \omega_0 T_s m = \frac{2\pi}{M}$
- Orthogonality property of complex exponentials: $a_i$ and $a_j$ are orthogonal (and thus independent) for $0 < i \neq j \leq M$.

# Band-limited Signal Recovery:

## Nyquist Recovery

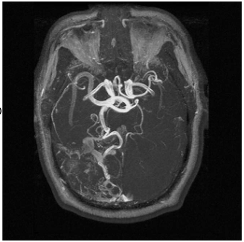


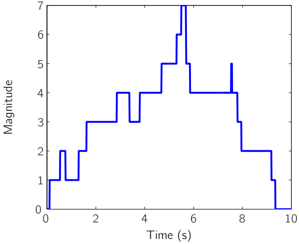

- Since rows are independent, need $M \geq 2N_b + 1$ to recover $x$.
- Implies $f_s \geq (2N_b + 1)\omega_0$: sampling frequency needs to be greater than two times bandwidth.

# So what is the problem?

- Signals often have high bandwidth, but lower complexity content

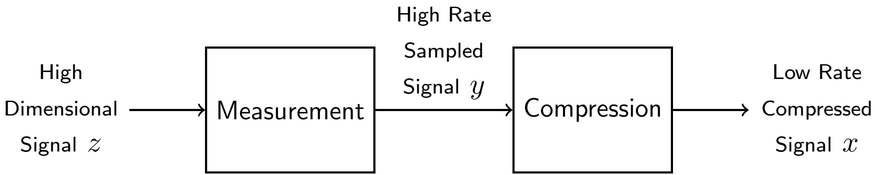
credit: J. Romberg






- What if we change the signal model: not bandlimited, but sparse in some basis.

# Current Solution: Measure Then Compress



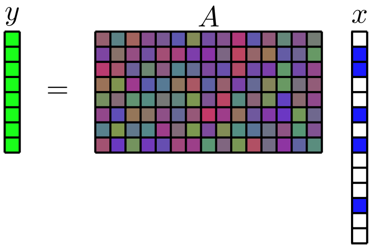

- Measurement costs \$. Compression costs \$.
- Can we combine the measurement and compression steps? (Compressive Sensing)

# Sparse Signal Recovery:

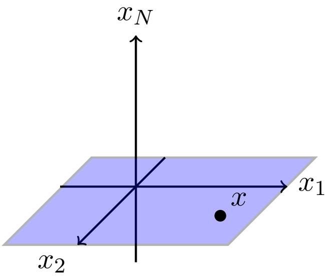
## Compressive Measurement Model

- Model: signal $x \in \mathbb{R}^N$, with $S$-sparse support, measurement, $y \in \mathbb{R}^M$.
  - $\Psi$ - signal basis (columns are $\psi_n$)
  - $\Phi$ - measurement matrix (rows are $\phi_m$)

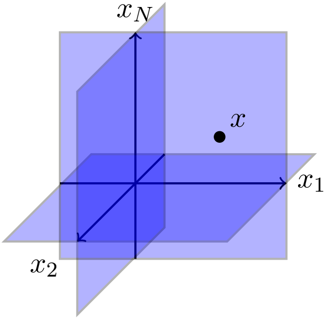
$$y = \underbrace{\Phi\Psi}_{A} x$$



# Geometry of Signal Models




Linear Subspace, dim $N_b$
Bandlimited Signals




Union of dim $S$ Subspaces
Sparse Signals

# Sparse Signal Recovery:

## Recovery via regularization

- Given $y$, can we recover $x$?
- $A$ is short and fat: non-trival null space means many solutions to $y = Ax$.
- Idea: regularized recovery

$$\hat{x} = \arg\min_{x} ||x||_{*} \quad \text{s.t. } y = Ax$$

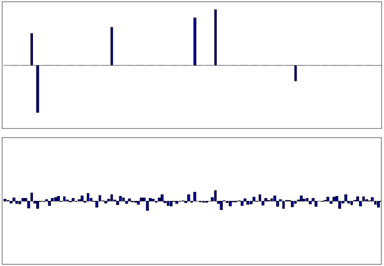
# Sparse Signal Recovery:

## $\ell_2$ recovery

- $\ell_2$-recovery (Euclidian distance) doesn't work

$$\hat{x} = \arg\min_x ||x||_2 \quad \text{s.t. } y = Ax$$

- Minimum is almost never sparse

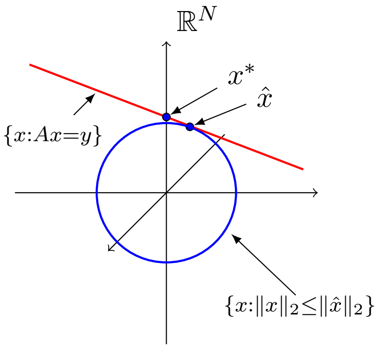


$x$

$\hat{x} = (A'A)^{-1}A'y$

# Sparse Signal Recovery:

## $\ell_2$ recovery geometry




Incorrect Recovery

# Sparse Signal Recovery:

## Sparcity preserving norms

- $\ell_0$-recovery: $\|x\|_0 = \#$ of non-zero elements of $x$.

$$\hat{x} = \arg\min_x ||x||_0 \quad \text{s.t. } y = Ax$$

  - Works generically if $M = S + 1$. However, computationally demanding.

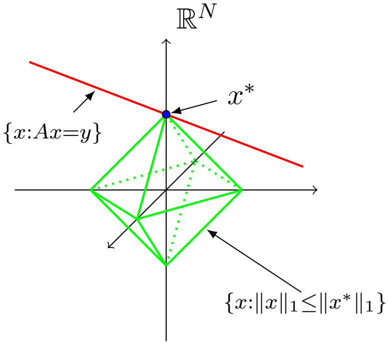
- $\ell_1$-recovery: $\|x\| = \sum |x_i|$. Convex! Recovery via LP:

$$\hat{x} = \arg\min_x ||x||_1 \quad \text{s.t. } y = Ax$$

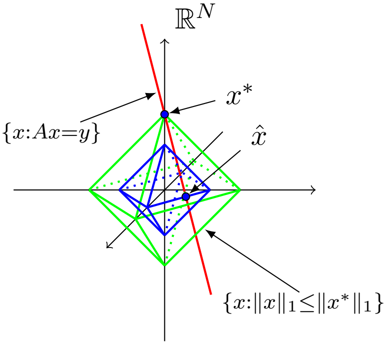
  - Also related to basis pursuit, lasso.
  - Works generically if $M \approx S \log N$!!!

# Sparse Signal Recovery:

## $\ell_1$ recovery geometry




Correct Recovery




Incorrect Recovery

# Other Recovery Methods

- Greedy methods - Orthogonal Matching Pursuit (Tropp, 2004)
- Iterative convex - Reweighted $\ell_1$ - (Candès, Wakin and Boyd, 2008)
- Non-convex - smoothed $\ell_0$ - (Chartrand, 2007; Mohimani et al., 2007)

# Recovery Example

256x256
original

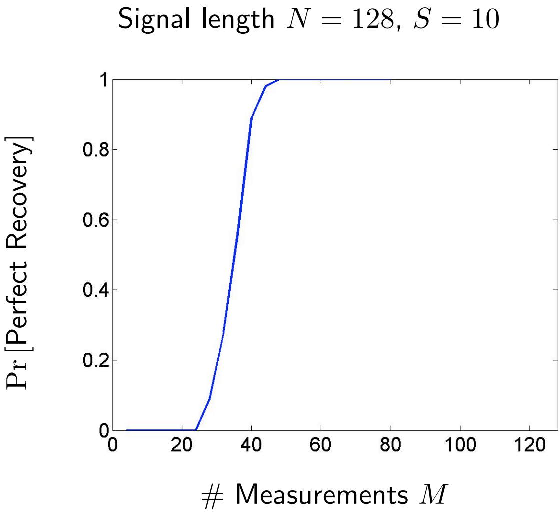
6500 wavelets

26000 random
projections

credit: J. Romberg

- Wavelets: 6500 largest coefficients
- 26000 random projections: recovery using wavelet basis
- Good approximation with 4x sampling rate over perfect knowledge

# Recovery Example



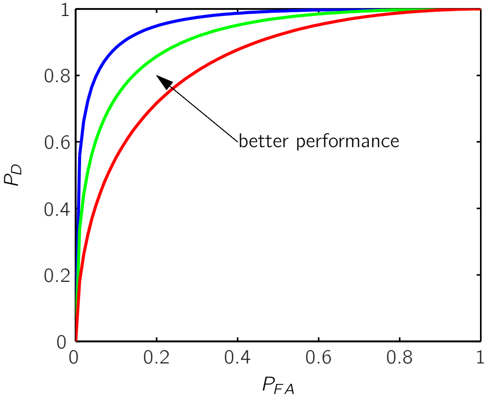
# Sparse Signal Detection

- The classic signal detection problem:
  - known signal $z$ may or may not have been sent
  - measurement $y$ corrupted by noise $v$
- Define events $\mathcal{E}_0$ and $\mathcal{E}_1$ as:

$$\mathcal{E}_0 \triangleq y = v$$
$$\mathcal{E}_1 \triangleq y = z + v$$

- Detection algorithm: decide if event $\mathcal{E}_0$ or $\mathcal{E}_1$ occurred.
- Performance metrics are
  - false-alarm probability - $P_{FA} = \Pr[(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_0)]$
  - detection probability - $P_D = \Pr[(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_1)]$

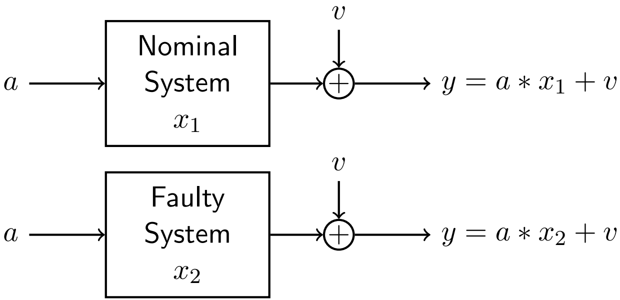
# Receiver Operation Characteristic: ROC curve




- How many measurements are necessary to obtain the desired performance?

# Fault Isolation

- Application: System with known fault condition. All signals are discrete time sequences.




- Subtract expected output: detection problem with $z = a * (x_1 - x_2)$
- Convolution: $z = A(x_1 - x_2)$, $A$ Toeplitz Matrix.

# Compressive Signal Processing

- Experiment model

$$y = \Phi\Psi x + v$$

  - $\Psi$ - signal basis (columns are $\psi_n$)
  - $\Phi$ - measurement matrix (rows are $\phi_m$)
  - $y \in \mathbb{R}^M$, measurement, $x \in \mathbb{R}^N$, $S$-sparse signal, $v \in \mathbb{R}^M$ measurement noise.
- Basic problems
  - Compressive recovery of unknown $S$-sparse signal using $M$ measurements, with $S < M \ll N$.
  - Detection of a known $S$-sparse signal using $M$ measurements, with $S < M \ll N$.

# Compressive Signal Processing:

## Questions

- What are the conditions that guarantee that all $x$ of a given sparsity can be recovered?
- What are the conditions that guarantee a particular level of performance in detection?
- How can we generate measurement matrices that meet these conditions?

# The Restricted Isometry Property (RIP)

- Introduced by Candès and Tao

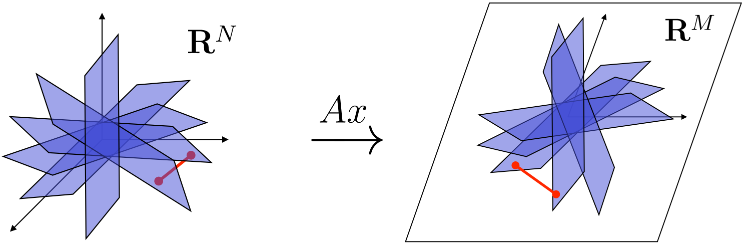
**Definition**

$X$ satisfies the RIP of order $S$ if there exists a $\delta_S \in (0,1)$ such that

$$(1-\delta_S)\,\|a\|_2^2 \le \|Xa\|_2^2 \le (1+\delta_S)\,\|a\|_2^2$$

holds for all $S$-sparse signals $a$.

# RIP as embedding




- Difference of two $S$-sparse signals is $2S$ sparse.

$$(1-\delta_{2S})\,\|u-v\|_2^2 \le \|A(u-v)\|_2^2 \le (1+\delta_{2S})\,\|u-v\|_2^2$$

# Recovery Result:

Candès (2008)

- Recovery algorithm (basis pursuit de-noising)

$$\hat{x} = \arg\min_x ||x||_1 \quad \text{s.t. } \|y - Ax\|_2 \le \epsilon$$

### Theorem

*Suppose $y$ is generated by $y = Ax^* + v$. If $A$ satisfies RIP with $\delta_{2S} < \sqrt{2} - 1$ and $\|v\|_2 < \epsilon$, then*

$$\|\hat{x} - x^*\|_2 \le C_0 \frac{\|x^* - x_s\|}{\sqrt{s}} + C_1 \epsilon$$

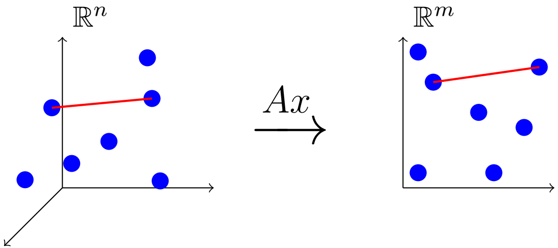
*where $x_s$ is the $S$-sparse approximation of $x^*$.*

- Implies perfect recovery if $x^*$ is $S$-sparse and no noise.

# Checking RIP

- Given $A$, does it satisfy RIP?
  - Check eigenvalues of each $M \times S$ submatrix - combinatorial.
- Generate $A$ *randomly* - satisfies RIP with high probability when $M = O(S \log N)$!
  - iid Gaussian entries
  - iid Bernoulli entries ($+$/- 1)
  - random Fourier ensemble
  - (Candes, Tao; Donoho; Traub, Wozniakowski; Litvak et al)
- Proofs bound eigenvalues of random matrices, but generally difficult to generalize to $\Psi \neq I$.

# Recall Johnson-Lindenstrauss Embedding




## J-L Embedding

Given $\epsilon > 0$ and set $\mathbb{P}$ of $P$ points in $\mathbb{R}^N$, find $A$ such that for all $u, v \in \mathbb{P}$,

$$(1-\epsilon)\|u-v\|^2 \leq \|A(u-v)\|^2 \leq (1+\epsilon)\|u-v\|^2$$

# Random J-L Embeddings

- Using our results from the last talk, we have the following:

**Theorem (Dasgupta and Gupta; Frankl; Achioptas; Indyk and Motwani)**

*Given set $\mathbb{P}$ of $P$ points in $\mathbb{R}^N$, choose $\epsilon > 0$ and $\beta > 0$. Let $A$ be an $M \times N$ matrix with independent elements $[A]_{ij} \sim \mathcal{N}\left(0, \frac{1}{M}\right)$ where*

$$M \geq \left(\frac{7 + 6\beta}{\min(.5, \epsilon)^2}\right) \ln(P).$$

*Then with probability greater than $1 - P^{-\beta}$, the following holds: For all $u, v \in \mathbb{P}$,*

$$(1 - \epsilon)\|u - v\|^2 \leq \|A(u - v)\|^2 \leq (1 + \epsilon)\|u - v\|^2$$

# Other Favorable Random Mappings:

## Sub-Gaussian Distributions

- In the proof, we used

$$[A]_{ij} \sim \mathcal{N}\left(0, \frac{1}{M}\right)$$

- Key step was Chernoff bound using moment generating function

### Definition

A random variable $X$ is *Sub-Gaussian* if there exists an $a \geq 0$ such that

$$\mathbb{E}\left[e^{sX}\right] \leq e^{\frac{a^2 s^2}{2}}$$

and $\tau$, the smallest such $a$, is called the *Gaussian standard* of $X$.

# Other Favorable Random Mappings:

## Properties of Sub-Gaussians

### Key Properties

- If $X_i$ are iid sub-Gaussian, $Y = \sum X_i$ is sub-Gaussian with standard $\tau_y \leq \sum \tau_{x_i}$
- If $X$ is sub-Gaussian with standard $\tau$, $\mathbb{E}\left[e^{sX^2}\right] \leq \frac{1}{1-2s\tau^2}$

# Other Favorable Random Mappings:

## Sub-Gaussian Examples

- We can use any zero mean sub-Gaussian iid sequence with variance $1/M$.
- Rademacher Sequence

$$[A]_{ij} = \begin{cases} +\frac{1}{\sqrt{M}} & \text{with probability } \frac{1}{2} \\ -\frac{1}{\sqrt{M}} & \text{with probability } \frac{1}{2} \end{cases}$$

- "Database-friendly" (Achlioptas)

$$[A]_{ij} = \begin{cases} +\sqrt{\frac{3}{M}} & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{1}{3} \\ -\sqrt{\frac{3}{M}} & \text{with probability } \frac{1}{6} \end{cases}$$

# From JL to RIP

- Baraniuk et al. (2008)
- Consider measurement with $\Psi = I$, $\Phi$ random elements from a favorable distribution

$$y = \underbrace{\Phi\Psi}_{A} x$$

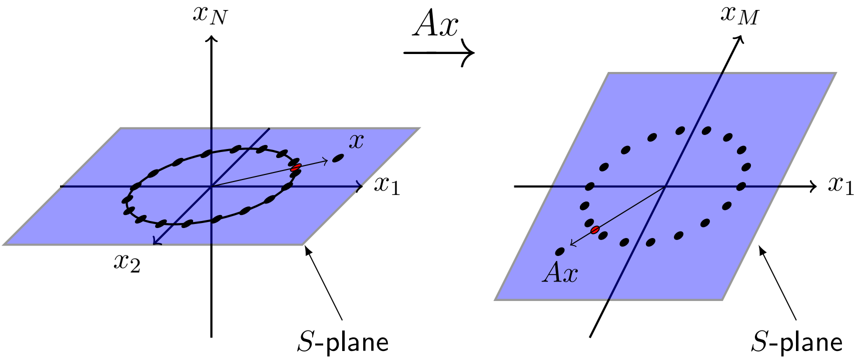
- Favorable distribution implies that for given $x \in \mathbb{R}^N$,

$$\Pr\left[\left|\|Ax\|_2^2 - \|x\|_2^2\right| \geq \epsilon\|x\|_2^2\right] \leq 2e^{-Mc_0(\epsilon)}$$

- pick $\epsilon = \delta_{2S}/2$

# From JL to RIP

- Examine mapping on one of $\binom{N}{S}$ $S$-planes in sparse model
  - Construct (careful) covering of unit sphere using $(12/\delta_{2S})^S$ points
  - JL: isometry for each point with high probability
  - Union bound for all points
  - Extend isometry to all $x$ in unit ball (and thus all $x$ in $S$-plane)



# A look at the probabilities:

## Union Bounds

- Probability of error $> \frac{\delta_{2S}}{2}$ when mapping 1 point

$$\leq 2e^{-Mc_0(\delta_{2S}/2)}$$

- Probability of error when $(12/\delta_{2S})^S$ points mapped

$$\leq 2(12/\delta_{2S})^S e^{-Mc_0(\delta_{2S}/2)}$$

- "Careful" covering implies that for all $x$ in unit ball, $\exists q$ in covering s.t. $\|x - q\| < \delta_{2S}/4$.
- Probability of error $> \delta_{2S}$ when unit ball mapped

$$\leq 2(12/\delta_{2S})^S e^{-Mc_0(\delta_{2S}/2)}$$

# A look at the probabilities, continued

- Probability of error $> \delta_{2S}$ when $\binom{N}{S}$ planes mapped:

$$\leq 2\binom{N}{S}(12/\delta_{2S})^k e^{-Mc_0(\delta_{2S}/2)} \leq 2e^{-c_0(\delta_{2S})M+S[\ln(eN/S)+\ln(12/\delta_{2S})]}$$

**Result**

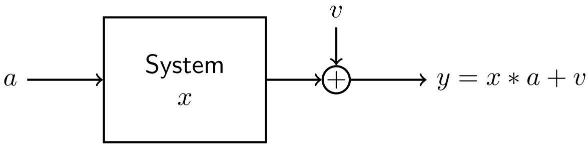
If $M > O(S\log(N/S))$, with probability greater than $1-2e^{-c_2 M}$, A random matrix with favorable distribution satisfies RIP.

- Bonus: Universality for orthonormal basis $\Psi$: only changes orientation of planes in model.

# Structured Measurements:

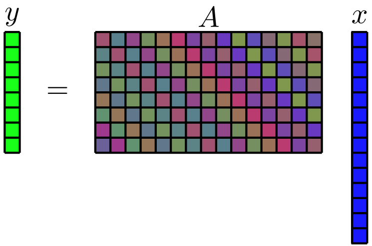
## A Detection Problem with Convolution

- We are not always free to choose the elements of $\Phi$ independently
  - Distributed measurements
  - Dynamic Systems



# Convolution implies Toeplitz measurement matrix

$$y = a * x$$




- Cannot choose the elements of $A$ independently

# Concentration of Measure for Toeplitz matrices

- Suppose $a$ is chosen iid Gaussian $x_i \sim \mathcal{N}\left(0, \frac{1}{M}\right)$.
- For fixed $x$, $y \sim \mathcal{N}\left(0, \frac{1}{M}P\right)$ where

$$[P]_{ij} = \sum_{i=1}^{n-|i-j|} x_i x_{i+|i-j|}$$

- Let $\rho(x) = \frac{\lambda_{\max}(P)}{\|x\|_2^2}$ and $\mu(x) = \frac{\frac{1}{d}\sum \lambda_i^2(P)}{\|x\|_2^2}$.

## Result

For any $\epsilon \in (0, 0.5)$

$$\Pr\left[\|Ax\|_2^2 \geq \|x\|_2^2 (1+\epsilon)\right] \leq e^{-\epsilon^2 M/6\rho(a)}$$

$$\Pr\left[\|Ax\|_2^2 \leq \|x\|_2^2 (1-\epsilon)\right] \leq e^{-\epsilon^2 M/4\mu(a)}$$

Sanandaji, Vincent, Wakin - 2010 Conference on Decision and Control

# Implications

- Result for $A$ Toeplitz:

$$\Pr\left[\|Ax\|_2^2 \geq \|x\|_2^2\,(1+\epsilon)\right] \leq e^{-\epsilon^2 M/6\rho(a)}$$
$$\Pr\left[\|Ax\|_2^2 \leq \|x\|_2^2\,(1-\epsilon)\right] \leq e^{-\epsilon^2 M/4\mu(a)}$$

- Recall result from previous lecture for $A$ unstructured:

$$\Pr\left[\|Ax\|_2^2 \geq \|x\|_2^2\,(1+\epsilon)\right] \leq e^{-\epsilon^2 M/6}$$
$$\Pr\left[\|Ax\|_2^2 \leq \|x\|_2^2\,(1-\epsilon)\right] \leq e^{-\epsilon^2 M/4}$$

- Concentration bound worsens over i.i.d. entries by factors $\rho$ and $\mu$.
- Bound: $\mu(a) \leq \rho(a) \leq \|a\|_0$. However, most $a$ are must less than this bound.

# Fault Detection Problem

- System impulse response can be $x_1$ or $x_2$.
- record $\tilde{y} = y - Ax_1$, let $\delta x = x_2 - x_1$
- Define events $\mathcal{E}_0$ and $\mathcal{E}_1$ as:

$$\mathcal{E}_0 \triangleq \tilde{y} = v$$
$$\mathcal{E}_1 \triangleq \tilde{y} = A\delta x + v$$

- Detection algorithm: decide if event $\mathcal{E}_0$ or $\mathcal{E}_1$ occurred.
- Performance metrics are
  - false-alarm probability - $P_{FA} = \Pr\left[(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_0)\right]$
  - detection probability - $P_D = \Pr\left[(\mathcal{E}_1 \text{ chosen when } \mathcal{E}_1)\right]$

# Neyman-Pearson Test

- The Neyman-Pearson detector maximizes $P_D$ for a given limit on failure probability, $P_{FA} \leq \alpha$ under Gaussian noise assumption.

$$\tilde{y}' A x \gtrless_{\mathcal{E}_0}^{\mathcal{E}_1} \gamma$$

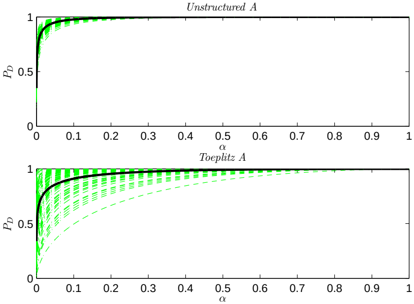
- Performance:

$$P_D = Q\left(Q^{-1}(P_{FA}) - \frac{\|Ax\|_2}{\sigma}\right)$$

- Since performance depends on $\|Ax\|_2$, worse performance for signals with large $\rho(a)$, $\mu(a)$.

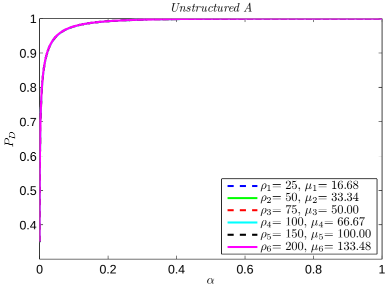
# Detection Performance

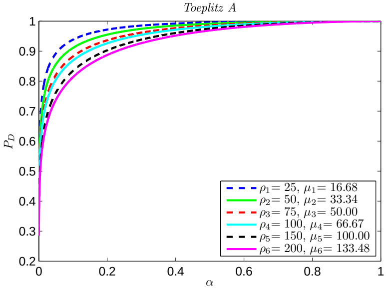
- $y = Ax + v$
- $A$ is $125 \times 250$
- $x$ is block sparse, $\mu(a) = 33$, $\rho(a) = 50$.
- Two cases:
    - $A$ - Unstructured
    - $A$ - Toeplitz
    - 1000 realizations of $A$



# Detection Performance

- Average detection performance for six different $x$.



# Conclusion

- Compressive Sensing - going beyond Nyquist sampling
- Sparse signal model with linear measurement model
- Recovery possible using convex optimization
- Work continues on
  - Recovery methods
  - Structured measurements
  - New applications - development of sparse signal models
  - ...