A Primer on Compressive Sensing

Phil Schniter

The Ohio State University
Duke University

Univ. of North Carolina, Chapel Hill
12/1/2016

Traditional sensing

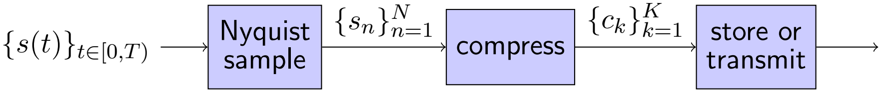
- We'd like to capture analog signals from the physical world and store them digitally on computers for subsequent processing, transmission, or reconstruction.
- Examples of "signals" include
  - speech or audio waveforms,
  - images (i.e., 2D waveforms),
  - video (i.e., 3D waveforms).
- The Nyquist theorem says that any bandlimited (i.e., smooth) signal can be sampled (giving a sequence) and then perfectly reconstructed.
- The Nyquist rate is the minimum sampling rate (i.e., # samples per unit time) needed for perfect reconstruction.

Traditional compression

Some signals are intrinsically simple and thus can be compressed without much loss of quality.

- Audio: MP3 gives roughly 10:1 compression relative to CD (=Nyquist)
- Images: JPEG gives roughly 25:1 compression relative to Nyquist
- Videos: MPEG gives roughly 100:1 compression relative to Nyquist

Compression facilitates efficient storage or transmission:



Compressive sensing

- Sometimes Nyquist sampling is too expensive.
- For compressible signals, Nyquist sampling is overkill.
- Can we do "compressive" sampling? Yes!
- Typical ingredients are:
    - randomly designed linear measurements
    - sparse signal representation
    - sophisticated signal reconstruction

Motivation

In some applications, measurements are costly:

- Magnetic resonance imaging:
  - scan time $\approx$ 30 minutes
  - scan time proportional to # samples taken

- Imaging outside visible spectrum:
  - CMOS doesn't work
  - high cost per pixel

- Wireless communication:
  - pilots inserted to measure channel
  - more pilots means less payload

System architecture

- Classical approach:

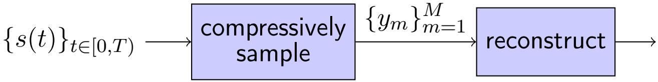
$$\{s(t)\}_{t\in[0,T)} \longrightarrow \boxed{\text{Nyquist sample}} \xrightarrow{\{s_n\}_{n=1}^{N}} \boxed{\text{compress}} \xrightarrow{\{c_k\}_{k=1}^{K}} \boxed{\text{reconstruct}} \longrightarrow$$

- New approach:

$$\{s(t)\}_{t\in[0,T)} \longrightarrow \boxed{\text{compressively sample}} \xrightarrow{\{y_m\}_{m=1}^{M}} \boxed{\text{reconstruct}} \longrightarrow$$

Nyquist rate $\frac{N}{T}$ $\gg$ compressive sampling rate $\frac{M}{T}$ $\gtrsim$ information rate $\frac{K}{T}$

Principal challenges in compressive sensing



1. Design of the compressive-measurement scheme

2. Reconstruction from the compressed measurements
   - We focus on recovering the Nyquist-rate signal samples $\{s_n\}_{n=1}^N$
   - Could easily reconstruct analog $\{s(t)\}_{t\in[0,T)}$ from Nyquist samples.

Simplifying assumptions

1. For now, assume noiseless linear measurements, e.g.,

$$y_m = \int_0^T \phi_m(t)\, s(t)\, dt, \quad m = 1, \dots, M$$

2. Also assume signal $s(t)$ is bandlimited, in which case Nyquist says

$$s(t) = \sum_{n=1}^{N} s_n \operatorname{sinc}\left(\frac{t}{T_s} - n + 1\right), \quad t \in [0, T).$$

Putting these together, we get the convenient discrete representation

$$y_m = \sum_{n=1}^{N} s_n \underbrace{\int_0^T \phi_m(t) \operatorname{sinc}\left(\frac{t}{T_s} - n + 1\right) dt}_{\triangleq\, \Phi_{m,n}}$$

or, in matrix/vector form, $\boxed{\boldsymbol{y} = \boldsymbol{\Phi} \boldsymbol{s}}$ for $\boldsymbol{s} \in \mathbb{R}^N$ and $\boldsymbol{y} \in \mathbb{R}^M$.

Design of linear measurements

Goal: design the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ so that

1. any signal $\boldsymbol{s}$ in class $\mathcal{S}$ can be reconstructed from $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{s}$,
2. the number of measurements M is minimal.

Key challenge:

There are fewer measurements M than unknowns N.

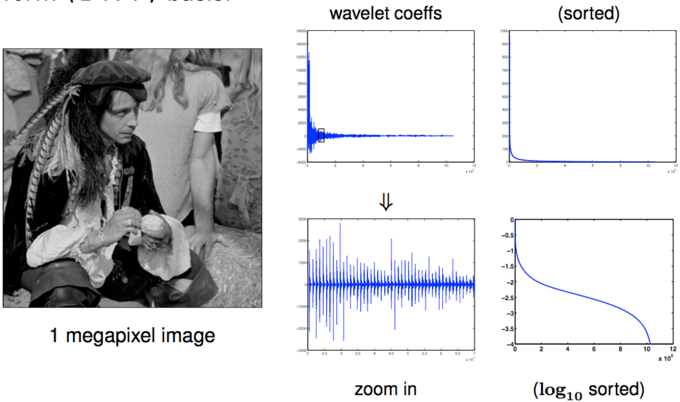
$\Rightarrow$ Many $\boldsymbol{s}$ satisfy the equation $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{s}$. How to find the correct $\boldsymbol{s}$?

Solution:

- If the signals in class $\mathcal{S}$ are sufficiently structured, only one of the $\boldsymbol{s}$ satisfying "$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{s}$" will be valid!
- Examples of structured signals include sparse signals, signals on manifolds, signals that can be expressed as low-rank matrices, etc.

Sparsity

- Many real-world signals are approximately sparse in a known basis.
- For example, natural images are sparse in the discrete wavelet transform (DWT) basis:



Typically: 99% signal energy captured by only 2.5% of DWT coefficients!

K-sparse in the dictionary $\boldsymbol{\Psi}$

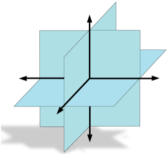
- We say that a signal class $\mathcal{S}$ is K-sparse in the dictionary $\boldsymbol{\Psi}$ if each $\boldsymbol{s} \in \mathcal{S}$ can be written as

$$\boldsymbol{s} = \boldsymbol{\Psi}\boldsymbol{x}$$

  for some K-sparse vector $\boldsymbol{x}$ (i.e., $\boldsymbol{x}$ has at most K nonzero elements).

- Usually orthonormal dictionaries $\boldsymbol{\Psi}$ are used (e.g., DWT, DCT, DFT), but overcomplete dictionaries may also be considered.

- Geometrically, a K-sparse vector $\boldsymbol{x} \in \mathbb{R}^N$ lives in a union of $\binom{N}{K}$ subspaces, each of dimension K:



Merging sparsity with linear compression

Recall...

- Linear measurement model: $\boldsymbol{y} = \boldsymbol{\Phi s}$ for $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$
- Sparse signal model: $\boldsymbol{s} = \boldsymbol{\Psi x}$ for K-sparse $\boldsymbol{x} \in \mathbb{R}^N$

Together...

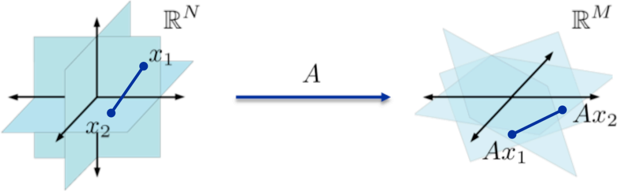
- Compressive sensing model: $\boldsymbol{y} = \underbrace{\boldsymbol{\Phi\Psi}}_{\triangleq \boldsymbol{A}} \boldsymbol{x}$ for $\boldsymbol{A} \in \mathbb{R}^{M \times N}$

Questions:

1. What properties of $\boldsymbol{A}$ ensure the recovery of $\boldsymbol{x}$?
2. Given dictionary $\boldsymbol{\Psi}$, how can we design $\boldsymbol{\Phi}$ to ensure agood $\boldsymbol{A}$?

Restricted isometry property

- Recall model: $\boxed{\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \text{ for } \boldsymbol{A} \in \mathbb{R}^{M\times N} \text{ and } K\text{-sparse } \boldsymbol{x} \in \mathbb{R}^N.}$
- Note: if signals $\boldsymbol{x}_1 \neq \boldsymbol{x}_2$ map to the same $\boldsymbol{y}$, they can't be recovered!



- In general, for our measurement system to be information preserving, we want that $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \approx \|\boldsymbol{A}\boldsymbol{x}_1 - \boldsymbol{A}\boldsymbol{x}_2\|_2$ for all K-sparse $\boldsymbol{x}_1, \boldsymbol{x}_2$, or

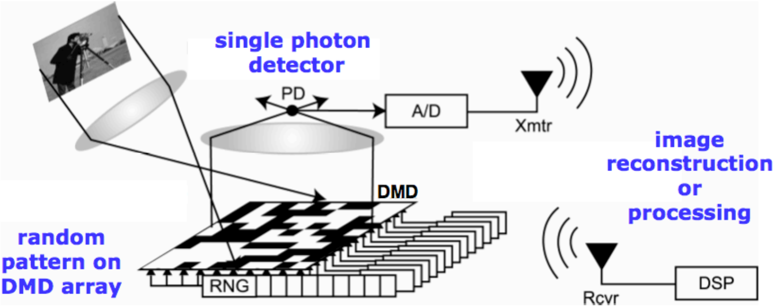
$$1 - \delta \leq \frac{\|\boldsymbol{A}\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_2^2} \leq 1 + \delta \quad \text{for all } 2K\text{-sparse } \boldsymbol{d}. \qquad \text{"RIP"}$$

Ensuring RIP with randomness

- Testing a given matrix for RIP is an NP-hard (combinatorial) problem.
- Fortunately, if $\boldsymbol{A}$ is randomly drawn with independent zero-mean sub-Gaussian entries (e.g., normal, ± 1), then *with high probability* it will satisfy RIP if
$$M \geq O\Big(K \log \frac{N}{K}\Big).$$
- Similarly, if $\boldsymbol{\Phi}$ is constructed randomly in the same way, then $\boldsymbol{A} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ will satisfy RIP for *any* orthonormal $\boldsymbol{\Psi}$.
- In practice, semi-random $\boldsymbol{\Phi}$ are preferable, e.g.,

  Create $\boldsymbol{\Phi} = \boldsymbol{J}\boldsymbol{F}\boldsymbol{D}$, where $\boldsymbol{D}$ is a diagonal matrix with random ± 1s, $\boldsymbol{F}$ is the N-FFT matrix, and $\boldsymbol{J}$ randomly selects M outputs.

Example: Single-pixel camera (Rice Univ.)



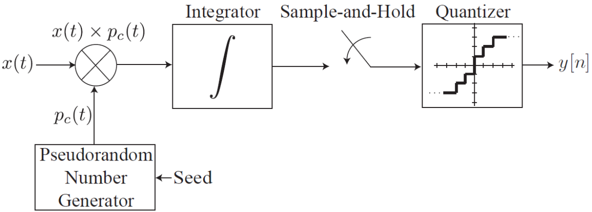
target
65536 pixels

11000 measurements
(16%)

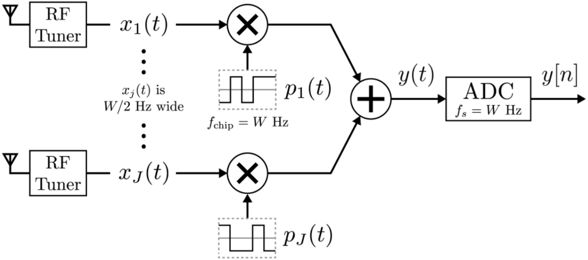
1300 measurements
(2%)

Other examples

Random demodulator:



Compressive multiplexor:



Signal recovery from compressive measurements

$$\{s(t)\}_{t\in[0,T)} \longrightarrow \boxed{\text{compressively sample}} \xrightarrow{\{y_m\}_{m=1}^{M}} \boxed{\text{reconstruct}} \longrightarrow$$

- So far we've talked about the design of the compressive sampler. Now we'll shift focus to signal reconstruction from compressed $\boldsymbol{y}$.

- In particular, we'll talk about how to reconstruct the Nyquist-rate signal samples $\boldsymbol{s}$ from

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\Phi s} + \boldsymbol{w} && \text{with additive measurement noise } \boldsymbol{w}! \\ &= \boldsymbol{\Phi \Psi x} + \boldsymbol{w} \\ &= \boldsymbol{Ax} + \boldsymbol{w} && \text{where } \boldsymbol{x} \text{ is approximately } K\text{-sparse} \end{aligned}$$

  In fact, recovering $\boldsymbol{x}$ is enough, since we can then construct $\boldsymbol{s} = \boldsymbol{\Psi x}$.

Sparse reconstruction

Goal: estimate $\boldsymbol{x} \in \mathbb{R}^N$ from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \in \mathbb{R}^M$ where

- $\boldsymbol{x}$ is approximately K-sparse (although K is unknown)
- $M \ll N$ but $M \geq K$
- $\boldsymbol{A}$ is RIP-like (all subsets of K columns are nearly orthonormal)

Popular methods:

- Convex methods based on ℓ_1-regularization
- Greedy search
- Bayesian inference

Best sparse fit — the ℓ_0 technique

Find the sparsest $\boldsymbol{x}$ that explains $\boldsymbol{y}$ up to a specified tolerance of ϵ:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \underbrace{\|\boldsymbol{x}\|_0}_{\#\text{ nonzero coefs}} \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

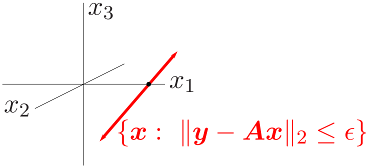
Unfortunately, this is NP-hard; we'd need to check all $\binom{N}{K} \approx N^K$ possible supports!

Let's think about this problem geometrically...

A toy example

Consider $\underset{\begin{bmatrix}\bullet\\\bullet\end{bmatrix}}{\boldsymbol{y}} = \underset{\begin{bmatrix}\bullet&\bullet&\bullet\\\bullet&\bullet&\bullet\end{bmatrix}}{\boldsymbol{A}} \underset{\begin{bmatrix}\bullet\\\bullet\\\bullet\end{bmatrix}}{\boldsymbol{x}} + \underset{\begin{bmatrix}\bullet\\\bullet\end{bmatrix}}{\boldsymbol{w}}$ with 1-sparse $\boldsymbol{x}$. $\begin{cases} M = 2 \\ N = 3 \\ K = 1 \end{cases}$

- The set of $\boldsymbol{x}$ such that $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon$ is described by an ϵ-thin rod.



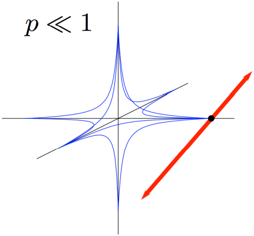
- The ℓ_0 technique would check increasingly large support hypotheses until it finds one whose signal subspace intersects the ϵ-rod. In this example, it would recover the true $\boldsymbol{x}$ if $\epsilon = 0$.

The geometry of constrained ℓ_p-minimization

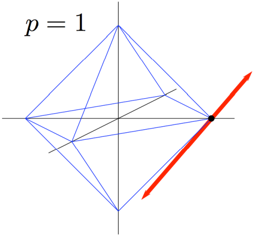
Now consider, for some fixed $p > 0$, the optimization problem:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \underbrace{\|\boldsymbol{x}\|_p}_{\sqrt[p]{\sum_n |x_n|^p}} \quad \text{s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon.$$

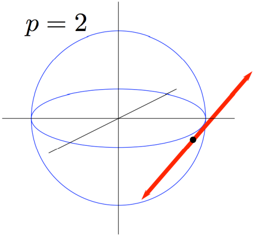
The solution can be found by growing the ℓ_p-ball until it touches the ϵ-rod:



Solution definitely sparse but problem is **NP hard**.



Solution usually sparse and problem is **convex**!



Solution is **not sparse**; $\Leftrightarrow$ LS when $\epsilon = 0$.

This suggests to use the ℓ_1 norm as a surrogate for the ℓ_0 norm!

LASSO

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \ \text{ s.t. } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2 \leq \epsilon$$

- Convex! Can be solved very efficiently.
- For $\boldsymbol{A}$ satisfying $2K$-RIP, LASSO guarantees that

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq \frac{C_1}{\sqrt{K}}\|\boldsymbol{x} - \boldsymbol{x}_K\|_1 + C_2\|\boldsymbol{w}\|_2$$

  where $\boldsymbol{x}_K$ is the best K-sparse approximation of $\boldsymbol{x}$ and C_1, C_2 are constants that depend on the RIP δ. Wow!
- In the special case when $\boldsymbol{x}$ is K-sparse, this simplifies to

$$\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq C_2\|\boldsymbol{w}\|_2.$$

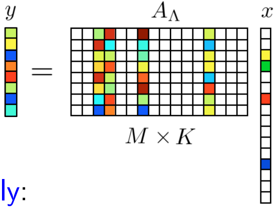
Greedy search

Main ideas:

- If we can correctly recover the support Λ of $\boldsymbol{x}$ (i.e., the locations of nonzeros), then determining the non-zero amplitudes is easy, e.g.,

$$\boldsymbol{x}_\Lambda = (\boldsymbol{A}_\Lambda^H \boldsymbol{A}_\Lambda)^{-1} \boldsymbol{A}_\Lambda^H \boldsymbol{y}$$

(least squares)



- Estimate the support sequentially:
  - Find the column of $\boldsymbol{A}$ most "similar" to $\boldsymbol{y}$ and store its index.
  - Subtract the effect of this column from $\boldsymbol{y}$.
  - Repeat (until residual is sufficiently small)!

Famous algorithms include MP, OMP, IHT, CoSaMP, Subspace Pursuit

Bayesian Methods

In the Bayesian approach, one . . .

- models the signal using a prior pdf $p(\boldsymbol{x})$,
- models the measurement process using a likelihood function $p(\boldsymbol{y}|\boldsymbol{x})$,
- performs inference via Bayes rule, yielding the posterior pdf

$$p(\boldsymbol{x}|\boldsymbol{y}) = Z^{-1}p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \quad \text{where } Z \text{ is a scaling constant,}$$

- often summarizes the posterior pdf by a point estimate like

$$\widehat{\boldsymbol{x}} = \int \boldsymbol{x}\, p(\boldsymbol{x}|\boldsymbol{y})\, d\boldsymbol{x} \qquad \text{MMSE estimate}$$

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}}\, p(\boldsymbol{x}|\boldsymbol{y}) \qquad \text{MAP estimate}$$

and possible other statistics that quantify estimate uncertainty.

Bayesian interpretation of LASSO

If we assume ...

- additive white Gaussian noise of variance σ^2
- i.i.d Laplacian signal with rate λ/σ^2

then

- likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{M/2}} \exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2))$
- prior: $p(\boldsymbol{x}) = \frac{1}{(2\sigma^2/\lambda)^M} \exp(-\frac{\lambda}{\sigma^2}\|\boldsymbol{x}\|_1)$

for which the maximum aposteriori (MAP) estimate is

$$\begin{aligned}\widehat{\boldsymbol{x}} &= \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} \log\left(Z^{-1} p(\boldsymbol{y}|\boldsymbol{x}) p(\boldsymbol{x})\right) \\ &= \arg\min_{\boldsymbol{x}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1\end{aligned}$$

which is an unconstrained version of the LASSO problem.

The relevance vector machine (RVM)

- The RVM is based on the *conditionally* Gaussian priors

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \prod_{n=1}^{N} \mathcal{N}(x_n; 0, \alpha_n^{-1}) \quad \text{and} \quad p(\boldsymbol{\alpha}) = \prod_{n=1}^{N} \Gamma(\alpha_n; 0, 0)$$
$$p(\boldsymbol{w}|\beta) \sim \prod_{m=1}^{M} \mathcal{N}(w_m; 0, \beta^{-1}) \quad \text{and} \quad \beta \sim \Gamma(0, 0)$$

  Note that, as "precision" $\alpha_n \to \infty$, the coefficient x_n is zeroed.

- The *conditional* posterior is (due to Gaussianity) simply

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}, \beta) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{for} \quad \begin{cases} \boldsymbol{\mu} = \beta\, \boldsymbol{\Sigma} \boldsymbol{A}^T \boldsymbol{y} \\ \boldsymbol{\Sigma} = \big(\beta \boldsymbol{A}^T \boldsymbol{A} + \mathcal{D}(\boldsymbol{\alpha})\big)^{-1}. \end{cases}$$

- In practice, $\{\boldsymbol{\alpha}, \beta\}$ are estimated using the EM algorithm and then plugged into $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to *approximate* the posterior $p(\boldsymbol{x}|\boldsymbol{y})$.

- The RVM (also known as "SBL" and "BCS") is relatively slow.

Other Bayesian methods

- Bayesian matching pursuits:
  - Greedy methods that use probabilistic support selection.

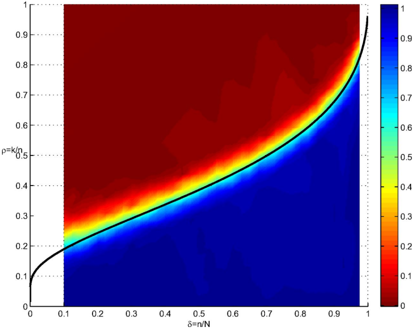
- Approximate message passing (AMP):
  - Inspired by methods from statistical physics and information theory.
  - Near-optimal in terms of speed and accuracy if $\boldsymbol{A}$ is large & random.

Phase transition curve (PTC) under large random $\boldsymbol{A}$

When examining a given algorithm's performance as a function of sampling ratio $\frac{M}{N}$ and sparsity ratio $\frac{K}{M}$, one finds a very sharp transition between perfect success and complete failure as $N, M, K \to \infty$.

LASSO:

$\frac{K}{M} \uparrow$

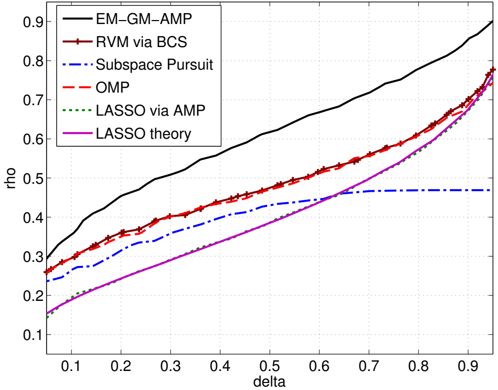


$\to \frac{M}{N}$

In some cases (e.g., LASSO), the PTC can be determined *analytically*.

Algorithm comparison 1

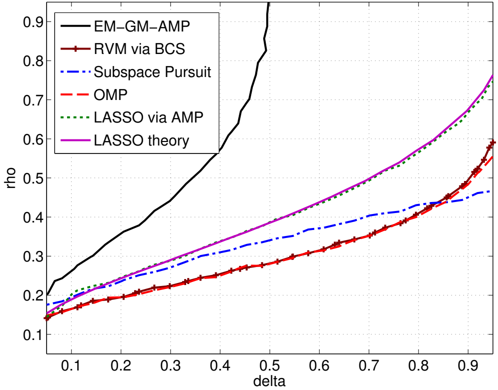
Recall: higher PTC = better algorithm.



Here, the non-zero elements of x were drawn independent zero-mean Gaussian.

Algorithm comparison 1

Recall: higher PTC = better algorithm.



Here, the non-zero elements of $\boldsymbol{x}$ were $= 1$.

More structure $\Rightarrow$ *possibility* for better performance.

Conclusions

Compressive sensing ...

- merges sampling and signal compression into a single operation
- is motivated by applications where cost-per-sample is high
- uses random linear measurements
- exploits the inherent sparsity of natural signals
- requires sophisticated algorithms for signal reconstruction.