Sparse Signal Recovery: Theory, Applications and Algorithms

Bhaskar Rao
Department of Electrical and Computer Engineering
University of California, San Diego

David Wipf
Biomedical Imaging Laboratory
University of California, San Francisco

Special Thanks to Yuzhe Jin
# Outline: Part 1

- Motivation for Tutorial
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - $\ell_1$ norm minimization
- Performance Guarantees
Outline: Part 2

- Motivation: Limitations of popular inverse methods
- *Maximum a posteriori* (MAP) estimation
- Bayesian Inference
- Analysis of Bayesian inference and connections with MAP
- Applications to neuroimaging
Outline: Part 1

- Motivation for Tutorial
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - $\ell_1$ norm minimization
- Performance Guarantees
Early Works

- Many More works
- Our first work
Early Session on Sparsity

Organized with Prof. Bresler a Special Session at the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPEC-DSP: SIGNAL PROCESSING WITH SPARSENESS CONSTRAINT</td>
<td></td>
</tr>
<tr>
<td>Signal Processing with the Sparseness Constraint</td>
<td>111-1861</td>
</tr>
<tr>
<td>B. Rao (University of California, San Diego, USA)</td>
<td></td>
</tr>
<tr>
<td>Application of Basis Pursuit in Spectrum Estimation</td>
<td>111-1865</td>
</tr>
<tr>
<td>S. Chen (IBM, USA); D. Donoho (Stanford University, USA)</td>
<td></td>
</tr>
<tr>
<td>Parsimony and Wavelet Method for Denoising</td>
<td>111-1869</td>
</tr>
<tr>
<td>H. Krim (MIT, USA); J. Pesquet (University Paris Sud, France); I. Schick (GTE Internetworking and Harvard Univ., USA)</td>
<td></td>
</tr>
<tr>
<td>Parsimonious Side Propagation</td>
<td>111-1873</td>
</tr>
<tr>
<td>P. Bradley, 0. Mangasarian (University of Wisconsin-Madison, USA)</td>
<td></td>
</tr>
<tr>
<td>Fast Optimal and Suboptimal Algorithms for Sparse Solutions to Linear Inverse Problems</td>
<td>111-1877</td>
</tr>
<tr>
<td>G. Harikumar (Tellabs Research, USA); C. Couvreur, Y. Bresler (University of Illinois, Urbana-Champaign, USA)</td>
<td></td>
</tr>
<tr>
<td>Measures and Algorithms for Best Basis Selection</td>
<td>111-1881</td>
</tr>
<tr>
<td>K. Kreutz-Delgado, B. Rao (University of California, San Diego, USA)</td>
<td></td>
</tr>
<tr>
<td>Sparse Inverse Solution Methods for Signal and Image Processing Applications</td>
<td>111-1885</td>
</tr>
<tr>
<td>B. Jeffs (Brigham Young University, USA)</td>
<td></td>
</tr>
<tr>
<td>Image Denoising Using Multiple Compaction Domains</td>
<td>111-1889</td>
</tr>
<tr>
<td>P. Ishwar, K. Ratakonda, P. Moulin, N. Ahuja (University of Illinois, Urbana-Champaign, USA)</td>
<td></td>
</tr>
</tbody>
</table>
Motivation for Tutorial

- Sparse Signal Recovery is an interesting area with many potential applications. Unification of the theory will provide synergy.

- Methods developed for solving the Sparse Signal Recovery problem can be a valuable tool for signal processing practitioners.

- Many interesting developments in the recent past that make the subject timely.
Outline: Part 1

- Motivation for Tutorial
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - $\ell_1$ norm minimization
- Performance Guarantees
Problem Description

- $y$ is $n \times 1$ measurement vector.
- $\Phi$ is $n \times m$ Dictionary matrix. $m >> n$.
- $x$ is $m \times 1$ desired vector which is sparse with $k$ non-zero entries.
- $\varepsilon$ is the additive noise modeled as additive white Gaussian.
Problem Statement

- **Noise Free Case**: Given a target signal $t$ and a dictionary $\Phi$, find the weights $x$ that solve:

$$\min_x \sum_{i=1}^{m} I(x_i \neq 0) \text{ subject to } y = \Phi x$$

where $I(.)$ is the indicator function.

- **Noisy Case**: Given a target signal $y$ and a dictionary $\Phi$, find the weights $x$ that solve:

$$\min_x \sum_{i=1}^{m} I(x_i \neq 0) \text{ subject to } \|y - \Phi x\|_2 \leq \beta$$
Complexity

- Search over all possible subsets, which would mean a search over a total of $\binom{m}{k}$ subsets. Combinatorial Complexity.

  With $m = 30; n = 20;$ and $k = 10$ there are $3 \times 10^7$ subsets (Very Complex)

- A branch and bound algorithm can be used to find the optimal solution. The space of subsets searched is pruned but the search may still be very complex.

- Indicator function not continuous and so not amenable to standard optimization tools.

**Challenge**: Find low complexity methods with acceptable performance
Outline: Part 1

- Motivation for Tutorial
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - $\ell_1$ norm minimization
- Performance Guarantees
Applications

- **Signal Representation** (Mallat, Coifman, Wickerhauser, Donoho, ...)
- **EEG/MEG** (Leahy, Gorodnitsky, Ioannides, ...)
- **Functional Approximation and Neural Networks** (Chen, Natarajan, Cun, Hassibi, ...)
- **Bandlimited extrapolations and spectral estimation** (Papoulis, Lee, Cabrera, Parks, ...)
- **Speech Coding** (Ozawa, Ono, Kroon, Atal, ...)
- **Sparse channel equalization** (Fevrier, Greenstein, Proakis, ...)
- **Compressive Sampling** (Donoho, Candes, Tao...)
- **Magnetic Resonance Imaging** (Lustig,..)
DFT Example

- **Measurement y**
  \[ y[l] = 2(\cos \omega_0 l + \cos \omega_1 l), \quad l = 0, 1, 2, ..., n-1. \quad n = 64. \]
  \[ \omega_0 = \frac{2\pi}{64} \frac{33}{2}, \quad \omega_1 = \frac{2\pi}{64} \frac{34}{2}. \]

- **Dictionary Elements:**
  \[ \phi_l^{(m)} = [1, e^{-j\omega_0}, e^{-j2\omega_0}, ..., e^{-j(n-1)\omega_0}]^T, \quad \omega_l = \frac{2\pi}{m} l \]

- Consider \( m = 64, 128, 256 \) and \( 512 \).

**Questions:**

- What is the result of a zero padded DFT?
- When viewed as problem of solving a linear system of equations dictionary, what solution does the DFT give us?
- Are there more desirable solutions for this problem?
DFT Example

- Note that

\[ y = \phi_{33}^{(128)} + \phi_{34}^{(128)} + \phi_{94}^{(128)} + \phi_{95}^{(128)} = \phi_{66}^{(256)} + \phi_{68}^{(256)} + \phi_{188}^{(256)} + \phi_{190}^{(256)} = \phi_{132}^{(512)} + \phi_{136}^{(512)} + \phi_{376}^{(512)} + \phi_{380}^{(512)} \]

- Consider the linear system of equations

\[ y = \Phi^{(m)} x \]

- The frequency components in the data are in the dictionaries \( \Phi^{(m)} \) for \( m = 128, 256, 512 \).

- What solution among all possible solutions does the DFT compute?
DFT Example
Sparse Channel Estimation

\[
r(i) = \sum_{j=0}^{m-1} s(i-j)c(j) - \epsilon(i), \quad i = 0, 1, ..., n-1
\]
Example: Sparse Channel Estimation

- Formulated as a sparse signal recovery problem

\[
\begin{bmatrix}
    r(0) \\
r(1) \\
    \vdots \\
r(n-1)
\end{bmatrix}
= \begin{bmatrix}
    s(0) & s(-1) & \cdots & s(-m+1) \\
    s(1) & s(0) & \cdots & s(-m+2) \\
    \vdots & \vdots & \ddots & \vdots \\
    s(n-1) & s(n-2) & \cdots & s(-m+n)
\end{bmatrix}
\begin{bmatrix}
    c(0) \\
c(1) \\
    \vdots \\
c(m-1)
\end{bmatrix}
+ \begin{bmatrix}
    \epsilon(0) \\
    \epsilon(1) \\
    \vdots \\
    \epsilon(n-1)
\end{bmatrix}
\]

- Can use any relevant algorithm to estimate the sparse channel coefficients
Compressive Sampling


Compressive Sampling

- Transform Coding

What is the problem here?
- Sampling at the Nyquist rate
- Keeping only a small amount of nonzero coefficients
- Can we directly acquire the signal below the Nyquist rate?
Compressive Sampling

- Transform Coding

- Compressive Sampling
Compressive Sampling

- Compressive Sampling

```
A       Ψ       x
\text{\Phi} \rightarrow

\text{Compute:}
1. Solve for \( w \) such that \( \Phi x = y \)
2. Reconstruction: \( b = \Psi x \)
```

- Issues
  - Need to recover sparse signal \( w \) with constraint \( \Phi x = y \)
  - Need to design sampling matrix \( A \)
Outline: Part 1

- Motivation for Tutorial
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - $\ell_1$ norm minimization
- Performance Guarantees
Potential Approaches

Combinatorial Complexity and so need alternate strategies

- **Greedy Search Techniques**: Matching Pursuit, Orthogonal Matching Pursuit

- **Minimizing Diversity Measures**: Indicator function not continuous. Define Surrogate Cost functions that are more tractable and whose minimization leads to sparse solutions, e.g. \( \ell_1 \) minimization

- **Bayesian Methods**: Make appropriate Statistical assumptions on the solution and apply estimation techniques to identify the desired sparse solution
Greedy Search Method: Matching Pursuit

- Select a column that is most aligned with the current residual

\[
y = \Phi x + \epsilon
\]

- \( r^{(o)} = y \)
- \( S^{(i)}: \) set of indices selected
- \( l = \arg \max_{1 \leq j \leq m} |\phi_j^T r^{(i-1)}| \)

- Remove its contribution from the residual
  - Update \( S^{(i)}: \) If \( l \not\in S^{(i-1)} \), \( S^{(i)} = S^{(i-1)} \cup \{l\} \). Or, keep \( S^{(i)} \) the same
  - Update \( r^{(i)}: \) \( r^{(i)} = P_{\phi_l} r^{(i-1)} = r^{(i-1)} - \phi_i \phi_i^T r^{(i-1)} \)

Practical stop criteria:
- Certain # iterations
- \( \| r^{(i)} \|_2 \) smaller than threshold
**Greedy Search Method: Orthogonal Matching Pursuit (OMP)**

- Select a column that is most aligned with the current residual

\[ y = \Phi x + \varepsilon \]

- \( r^{(0)} = y \)
- \( S^{(i)} \): set of indices selected
- \( l = \arg\max_{1 \leq j \leq m} \phi_j^T r^{(i-1)} \)

- Remove its contribution from the residual
  - Update \( S^{(i)} \): \( S^{(i)} = S^{(i-1)} \cup \{l\} \)
  - Update \( r^{(i)} \): \( r^{(i)} = P_{[l_1, l_2, \ldots, l_i]} r^{(i-1)} = r^{(i-1)} - P_{[l_1, l_2, \ldots, l_i]} r^{(i-1)} \)
Greedy Search Method: Order Recursive OMP

- Select a column that is most aligned with the current residual

\[ y = \Phi x + \varepsilon \]

- \( r^{(0)} = y \)
- \( S^{(i)}: \) set of indices selected
- \( l = \arg \max_{1 \leq j \leq m} |\phi_j^T r^{(i-1)}|^2 / \|\phi_j^{(i-1)}\|_2^2 \)

- Remove its contribution from the residual
  - Update \( S^{(i)}: S^{(i)} = S^{(i-1)} \cup \{l\} \)
  - Update \( r^{(i)}: r^{(i)} = P_{[l_1, l_2, \ldots, l_i]} r^{(i-1)} = r^{(i-1)} - P_{[l_1, l_2, \ldots, l_i]} r^{(i-1)} \)
  - Update \( \|\phi_l^{(i)}\|_2^2: \phi_l^{(i)} = P_{[l_1, l_2, \ldots, l_i]} \phi_l \). Can be computed recursively
Deficiency of Matching Pursuit
Type Algorithms

• If the algorithm picks a wrong index at an iteration, there is no way to correct this error in subsequent iterations.

Some Recent Algorithms
• Stagewise Orthogonal Matching Pursuit (Donoho, Tsaig, ..)
• COSAMP (Needell, Tropp)
Inverse Techniques

- For the systems of equations $\Phi x = y$, the solution set is characterized by $\{x_s : x_s = \Phi^+ y + v; v \in N(\Phi)\}$, where $N(\Phi)$ denotes the null space of $\Phi$ and $\Phi^+ = \Phi^T(\Phi \Phi^T)^{-1}$.

- **Minimum Norm solution**: The minimum $\ell_2$ norm solution $x_{mn} = \Phi^+ y$ is a popular solution.

- **Noisy Case**: regularized $\ell_2$ norm solution often employed and is given by

$$x_{reg} = \Phi^T(\Phi \Phi^T + \lambda I)^{-1} y$$
Minimum 2-Norm Solution

- **Problem**: Minimum $\ell_2$ norm solution is not sparse

Example:

$$\Phi = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$x_{mn} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} & \frac{1}{3} \end{bmatrix}^T$$ vs. $$x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$$

DFT: Also computes minimum 2-norm solution
Diversity Measures

- Recall:

\[
\min_x \sum_{i=1}^{m} l(x_i \neq 0) \quad \text{subject to} \quad y = \Phi x
\]

- Functionals whose minimization leads to sparse solutions

- Many examples are found in the fields of economics, social science and information theory

- These functionals are usually concave which leads to difficult optimization problems
Examples of Diversity Measures

- $\ell_{(p \leq 1)}$ Diversity Measure

$$E^{(p)}(x) = \sum_{i=1}^{m} |x_i|^p, \ p \leq 1$$

- As $p \to 0$,

$$\lim_{p \to 0} E^{(p)}(x) = \lim_{p \to 0} \sum_{i=1}^{m} |x_i|^p = \sum_{i=1}^{m} I(x_i \neq 0)$$

- $\ell_1$ norm, convex relaxation of $\ell_0$

$$E^{(1)}(x) = \sum_{i=1}^{m} |x_i|$$
\( \ell_1 \) Diversity Measure

- **Noiseless case**

\[
\min_x \sum_{i=1}^m |x_i| \quad \text{subject to} \quad \Phi x = y
\]

- **Noisy case**
  - \( \ell_1 \) regularization [Candes, Romberg, Tao]

\[
\min_x \sum_{i=1}^m |x_i| \quad \text{subject to} \quad \|y - \Phi x\|_2 \leq \beta
\]

  - Lasso [Tibshirani], Basis Pursuit De-noising [Chen, Donoho, Saunders]

\[
\min_x \|y - \Phi x\|_2^2 + \lambda \sum_{i=1}^m |x_i|
\]
$\ell_1$ norm minimization and MAP estimation

- MAP estimation

$$\hat{x} = \arg\max_x p(x | y)$$

$$= \arg\max_x [\log p(y | x) + \log p(x)]$$

- If we assume
  - $\epsilon_i$ is zero mean, i.i.d. Gaussian noise
  - $p(x) = \prod_i p(x_i)$, where $p(x_i) \propto \exp(-\lambda |x_i|)$

- Then

$$\hat{x} = \arg\min_x \left[ \|y - \Phi x\|_2^2 + \lambda \sum_i |x_i| \right]$$
Attractiveness of $\ell_1$ methods

- Convex Optimization and associated with rich class of optimization algorithms
  - Interior-point methods
  - Coordinate descent method
  - .......

- Question
  - What is the ability to find the sparse solution?
Why diversity measure encourages sparse solutions?

$$\min \left\| [x_1, x_2]^T \right\|_p \quad \text{subject to} \quad \phi_1 x_1 + \phi_2 x_2 = y$$

$$\phi_1 x_1 + \phi_2 x_2 = y$$

$0 \leq p < 1$  
$p = 1$  
$p > 1$
$l_1$ norm and linear programming

Linear Program (LP): $\min_x c^T x$ subject to $\Phi x = y$

Key Result in LP (Luenberger):

a) If there is a feasible solution, there is a basic feasible solution*

b) If there is an optimal feasible solution, there is an optimal basic feasible solution.

* If $\Phi$ is $n \times m$, then a basic feasible solution is a solution with $n$ non-zero entries
Example with $\ell_1$ diversity measure

$\Phi = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

- **Noiseless Case**
  - $x_{BP} = [1, 0, 0]^{T}$  
    (machine precision)

- **Noisy Case**
  - Assume the measurement noise $\varepsilon = [0.01, -0.01]^{T}$
  - $\ell_1$ regularization result: $x_{\ell_1R} = [0.986, 0, 8.77 \times 10^{-6}]^{T}$
  - Lasso result ($\lambda = 0.05$): $x_{lasso} = [0.975, 0, 2.50 \times 10^{-5}]^{T}$
Example with $\ell_1$ diversity measure

- Continue with the DFT example:
  
  $$y[l] = 2(\cos \omega_0 l + \cos \omega_1 l), \ l = 0, 1, 2, ..., n-1. \ n = 64.$$  
  
  $$\omega_0 = \frac{2\pi \cdot 33}{64}, \ \omega_1 = \frac{2\pi \cdot 34}{64}.$$  

- 64, 128, 256, 512 DFT cannot separate the adjacent frequency components.

- Using $\ell_1$ diversity measure minimization (m=256)
Outline: Part 1

- Motivation for Tutorial
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - $\ell_1$ norm minimization
- Performance Guarantees
Important Questions

- When is the $\ell_0$ solution unique?
- When is the $\ell_1$ solution equivalent to that of $\ell_0$?
  - Noiseless Case
  - Noisy Measurements
Uniqueness

• Definition of Spark
  ◦ The smallest number of columns from $\Phi$ that are linearly dependent. ($\text{Spark}(\Phi) \leq n+1$)

• Uniqueness of sparsest solution
  ◦ If $\sum_{i=1}^{m} l(x_i \neq o) < \frac{1}{2} \text{Spark}(\Phi)$, then $x$ is the unique solution to

$$\text{argmin}_x \sum_{i=1}^{m} l(x_i \neq o) \text{ subject to } \Phi x = y$$
Mutual Coherence

- For a given matrix $\Phi = [\phi_1, \phi_2, \ldots, \phi_m]$, the mutual coherence $\mu(\Phi)$ is defined as

$$\mu(\Phi) \triangleq \max_{1 \leq i, j \leq m; i \neq j} \frac{\|\phi_i^T \phi_j\|}{\|\phi_i\|_2 \|\phi_j\|_2}$$
Performances of $\ell_1$ diversity minimization algorithms

- **Noiseless Case** [Donoho & Elad 03]

If
$$\sum_{i=1}^{m} I(x_i \neq 0) < \frac{1}{2} \left( 1 + \frac{1}{\mu(\Phi)} \right),$$

then $x$ is the unique solution to

$$\arg\min_{x} \sum_{i=1}^{m} |x_i| \quad \text{subject to} \quad \Phi x = y$$
Performances of $\ell_1$ diversity minimization algorithms

- **Noisy Case** [Donoho et al 06]

Assume $y = \Phi x + \varepsilon, \|\varepsilon\|_2 \leq \beta, \sum_{i=1}^{m} I(x_i \neq 0) < \frac{1}{4} \left( 1 + \frac{1}{\mu(\Phi)} \right)$

Then the solution

$$d^* = \arg\min_d \sum_{i=1}^{m} |d_i| \quad \text{subject to} \quad \|y - \Phi d\|_2 \leq \beta$$

satisfies

$$\|d^* - x\|_2^2 \leq \frac{4\beta^2}{1 - \mu(\Phi)(4\|x\|_0 - 1)}$$
Empirical distribution of mutual coherence (Gaussian matrices)

- Gaussian Matrices: \( N(0,1) \), normalize column norm to 1.
- 2000 random generated matrices \( \Phi \)
- Histogram of mutual coherence

\[
\Phi_{100 \times 200}, \text{ mean } = 0.4025, \text{ std } = 0.0249
\]

\[
\Phi_{1000 \times 2000}, \text{ mean } = 0.161, \text{ std } = 0.0073
\]
Performance of Orthogonal Matching Pursuit

- **Noiseless Case** [Tropp 04]

\[
\text{If } \sum_{i=1}^{m} I(x_i \neq 0) < \frac{1}{2} \left( 1 + \frac{1}{\mu(\Phi)} \right), \text{ then OMP guarantees recovery of } x \text{ after } \sum_{i=1}^{m} I(x_i \neq 0) \text{ iterations.}
\]
Performance of Orthogonal Matching Pursuit

- Noisy Case [Donoho et al]

Assume \( y = \Phi x + \varepsilon, \|\varepsilon\|_2 \leq \beta, \quad x_{\min} = \min_{1 \leq i \leq m} |x_i|, \)

and \( \sum_{i=1}^{m} I(x_i \neq 0) < \frac{1}{2} \left( 1 + \frac{1}{\mu(\Phi)} \right) - \frac{\beta}{\mu(\Phi) \cdot x_{\min}} \).

Stop OMP when residual error \( \leq \beta \).

Then the solution of OMP satisfies

\[
\|x_{\text{OMP}} - x\|_2^2 \leq \frac{\beta^2}{1 - \mu(\Phi)(\|x\|_0 - 1)}
\]
Restricted Isometry Constant

- Definition [Candes et al]

For a matrix $\Phi$, the smallest constant $\delta_k$ such that

$$(1 - \delta_k)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_k)\|x\|_2^2$$

for any $k$-sparse signal $x$. 
Performances of $\ell_1$ diversity measure minimization algorithms

- [Candes 08]

Assume $y = \Phi x + \varepsilon$, $x$ is $k$-sparse, $\|\varepsilon\|_2 \leq \beta$ and $\delta_{2k} < \sqrt{2} - 1$.

Then, the solution

$$d^* = \arg\min_d \sum_{i=1}^{m} |d_i| \text{ subject to } \|y - \Phi d\|_2 \leq \beta$$

satisfies

$$\|d^* - x\|_2 \leq C \cdot \beta$$

where $C$ only depends on $\delta_{2k}$. 
Performances of $\ell_1$ diversity measure minimization algorithms

- [Candes 08]

Assume $y = \Phi x + \varepsilon, \|\varepsilon\|_2 \leq \beta$ and $\delta_{2k} < \sqrt{2} - 1$.

Then, the solution

$$d^* = \arg\min_d \sum_{i=1}^m |d_i| \quad \text{subject to} \quad \|y - \Phi d\|_2 \leq \beta$$

satisfies

$$\|d^* - x\|_2 \leq C_1 \cdot \frac{1}{\sqrt{k}} \|x - x_k\|_1 + C_2 \cdot \beta$$

where $C_1, C_2$ only depend on $\delta_{2k}$ and $x_k$ is the vector $x$ with all but the $k$-largest entries set to zero.
Matrices with Good RIC

• It turns out that random matrices can satisfy the requirement (say, $\delta_{2k} < \sqrt{2-1}$) with high probability.

• For a matrix $\Phi_{n \times m}$
  ◦ Generate each element $\phi_{ij} \sim N(0, 1/n)$, i.i.d.
  ◦ [Candes et al] If $n = O\left(k \log\left(\frac{m}{k}\right)\right)$, then $P(\delta_{2k} < \sqrt{2-1}) \to 1$.

• Observations:
  ◦ A large Gaussian random matrix will have good RIC with high probability.
  ◦ Similar results can be obtained using other probability ensembles: Bernoulli, Random Fourier, ....

• For $\ell_1$ based procedure, number of measurements required are $n \geq k \log m$
More General Question

- What are limits of recovery in the presence of noise?
  - No constraint on recovery algorithm

- Information theoretic approaches are useful in this case (Wainwright, Fletcher, Akcakaya, ..)

- Can connect the problem of support recovery to channel coding over the Gaussian multiple access channel. Capacity regions become useful (Jin)
Performance for Support Recovery

- Noisy measurements: $y = \Phi x + \epsilon, \; \epsilon_i \sim N(0, \sigma_\epsilon^2)$, i.i.d.
- Random matrix $\Phi$, $\phi_{ij} \sim N(0, \sigma_\Phi^2)$, i.i.d.
- Performance metric: exact support recovery
- Consider a sequence of problems with increasing sizes
- $k = 2$: $c(x) \overset{\Delta}{=} \min_{T \subseteq \{1, 2\}} \left\{ \frac{1}{2|T|} \log \left( 1 + \frac{\sigma_\Phi^2}{\sigma_\epsilon^2} \sum_{i \in T} \chi_{\text{nonzero}, i}^2 \right) \right\}$. (Two-user MAC capacity)

**Sufficient condition**

If

$$\limsup_{m \to \infty} \frac{\log m}{n_m} < c(x)$$

then there exists a sequence of support recovery methods (for diff. problems resp.) such that

$$\lim_{m \to \infty} P\{\text{supp}(\hat{x}) \neq \text{supp}(x)\} = 0$$

**Necessary condition**

If there exists a sequence of support recovery methods such that

$$\lim_{m \to \infty} P\{\text{supp}(\hat{x}) \neq \text{supp}(x)\} = 0$$

then

$$\limsup_{m \to \infty} \frac{\log m}{n_m} \leq c(x)$$

- The approach can deal with general scaling among $(m, n, k)$. 

54
Network information theory perspective
Connection to channel coding ($K = 1$)

\[
y = \Phi x + \varepsilon
\]

this column is selected
\[ y = x_i \phi_i + \epsilon \]
Additive White Gaussian Noise (AWGN) Channel

- \( a \): channel input
- \( h \): channel gain
- \( e \): additive Gaussian noise
- \( b \): channel output, \( b = ha + e \)

Recall
\[
y = x_i \phi_i + \epsilon
\]
Connection to channel coding: \((K \geq 1)\)

\[
y = \Phi x + \varepsilon
\]
Multiple Access Channel (MAC)

\[ y = x_{s_1} \phi_{s_1} + \ldots + x_{s_k} \phi_{s_k} + \varepsilon \]

\[ b = h_1 a_1 + \ldots + h_K a_K + e \]
Connection between two problems

\[
\begin{align*}
\Phi : & \text{dictionary, measurement matrix} \quad \text{codebook} \\
\phi_i : & \text{a column} \quad \text{a codeword} \\
\Phi : & \text{dictionary, measurement matrix} \\
\phi_i & \text{a column} \quad \text{a codeword} \\
m \text{different positions in } x & \quad \text{message set } \{1, 2, \ldots, m\} \\
 s_{1}, \ldots, s_{k} : & \text{indices of nonzeros} \quad \text{the messages selected} \\
x_{s_{i}} : & \text{source activity} \quad \text{channel gain } h_i
\end{align*}
\]
### Differences between two problems

<table>
<thead>
<tr>
<th>Channel coding</th>
<th>Support recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>Each sender works with a codebook designed <strong>only</strong> for that sender.</td>
<td>All “senders” share the <strong>same</strong> codebook. Different senders select different codewords. <strong>(Common codebook)</strong></td>
</tr>
<tr>
<td>Capacity result available when channel gain $h_i$ is <strong>known</strong> at receiver.</td>
<td>We <strong>don’t</strong> know the nonzero signal activities $x_i$. <strong>(Unknown channel)</strong></td>
</tr>
</tbody>
</table>

- **Proposed approach** [Jin, Kim and Rao]
  - Leverage MAC capacity result to shed light on performance limit of exact support recovery.
  - Conventional methods + customized methods.
    - Distance decoding
    - Nonzero signal value estimation
    - Fano’s Inequality
Sparse Signal Recovery: Theory, Applications and Algorithms: Part II

Bhaskar Rao
Digital Signal Processing Lab
UC San Diego

David Wipf
Biomagnetic Imaging Lab
UC San Francisco
Outline

1. Motivation: Limitations of popular inverse methods
2. *Maximum a posteriori* (MAP) estimation
3. Bayesian Inference
4. Analysis of Bayesian inference and connections with MAP
5. Applications to neuroimaging
Section I: Motivation
Limitation I

- Most sparse recovery results, using either greedy methods (e.g., OMP) or convex $\ell_1$ minimization (e.g., BP), place heavy restrictions on the form of the dictionary $\Phi$.

- While in some situations we can satisfy these restrictions (e.g., compressive sensing), for many applications we cannot (e.g., source localization, feature selection).

- When the assumptions on the dictionary are violated, performance may be very suboptimal.
Limitation II

- The distribution of nonzero magnitudes in the maximally sparse solution $x_0$ can greatly affect the difficulty of canonical sparse recovery problems.

- This effect is strongly algorithm-dependent …
Examples:

$\ell_1$-norm Solutions and OMP

- With $\ell_1$-norm solutions, performance is independent of the magnitudes of nonzero coefficients [Malioutov et al., 2004].

- **Problem**: Performance does not improve when the situation is favorable.

- **OMP** is highly sensitive to these magnitudes.

- **Problem**: Perform degrades heavily with unit magnitudes.
If the magnitudes of the non-zero elements in $x_0$ are highly scaled, then the canonical sparse recovery problem should be easier.

The (approximate) Jeffreys distribution produces sufficiently scaled coefficients such that best solution can always be easily computed.
Jeffreys Distribution

Density: \( p(x) \propto \frac{1}{|x|} \)
Empirical Example

- For each test case:
  1. Generate a random dictionary $\Phi$ with 50 rows and 100 columns.
  2. Generate a sparse coefficient vector $\mathbf{x}_0$.
  3. Compute signal via $\mathbf{y} = \Phi \mathbf{x}_0$ (noiseless).
  4. Run $\textsc{BP}$ and $\textsc{OMP}$, as well as a competing Bayesian method called $\textsc{SBL}$ (more on this later) to try and correctly estimate $\mathbf{x}_0$.
  5. Average over 1000 trials to compute empirical probability of failure.

- Repeat with different sparsity values, i.e., $\|\mathbf{x}_0\|_0$ ranging from 10 to 30.
Sample Results \( (n = 50, m = 100) \)
Limitation III

- It is not immediately clear how to use these methods to assess uncertainty in coefficient estimates (e.g., covariances).

- Such estimates can be useful for designing an optimal (non-random) dictionary $\Phi$.

- For example, it is well known in and machine learning and image processing communities that under-sampled random projections of natural scenes are very suboptimal.
Section II:

MAP Estimation Using the Sparse Linear Model
Overview

♦ Can be viewed as sparse penalized regression with general sparsity-promoting penalties ($\ell_1$ penalty is special case).

♦ These penalties can be chosen to overcome some previous limitations when minimized with simple, efficient algorithms.

♦ Theory is somewhat harder to come by, but practical results are promising.

♦ In some cases can guarantee improvement over $\ell_1$ solution on sparse recovery problems.
Sparse Linear Model

- Linear generative model:

\[ y = \Phi x + \varepsilon \]

- **Objective**: Estimate the unknown \( x \) given the following assumptions:

1. \( \Phi \) is *overcomplete*, meaning the number of columns \( m \) is greater than the signal dimension \( n \).
2. \( x \) is *maximally sparse*, i.e., many elements equal zero.
Sparse Inverse Problem

- **Noiseless case ($\varepsilon = 0$):**
  \[
  x_0 \triangleq \arg \min_x \|x\|_0 \quad \text{s.t.} \quad y = \Phi x
  \]

- **Noisy case ($\varepsilon > 0$):**
  \[
  x_0(\lambda) \triangleq \arg \min_x \|y - \Phi x\|_2^2 + \lambda \|x\|_0
  = \arg \max_x \exp \left[ -\frac{1}{2\lambda} \|y - \Phi x\|_2^2 \right] \exp \left[ -\frac{1}{2} \|x\|_0 \right]
  \]

  \[
  \ell_0 \text{ norm } = \# \text{ of nonzeros in } x
  \]

  **likelihood** \quad **prior**
Difficulties

1. Combinatorial number of local minima

2. Objective is discontinuous

A variety of existing approximate methods can be viewed as MAP estimation using a flexible class of sparse priors.
Sparse Priors: 2-D Example

Gaussian Distribution

Sparse Distribution

[Tipping, 2001]
Basic MAP Estimation

\[ \hat{x} \triangleq \arg \max_x p(x \mid y) \]
\[ = \arg \min_x -\log p(y \mid x) - \log p(x) \]
\[ = \arg \min_x \| y - \Phi x \|^2_2 + \lambda \sum_{i=1}^{n} g(x_i) \]

\( g(x_i) = h(x_i^2) \),
where \( h \) is a nondecreasing, concave function

**Note**: Bayesian interpretation will be useful later …
Example Sparsity Penalties

♦ With \( g(x_i) = I[x_i \neq 0] \) we have the canonical sparsity penalty and its associated problems.

♦ Practical selections:

\[
\begin{align*}
g(x_i) &= \log(x_i^2 + \epsilon), \quad [\text{Chartrand and Yin, 2008}] \\
g(x_i) &= \log(|x_i| + \epsilon), \quad [\text{Candes et al., 2008}] \\
g(x_i) &= |x_i|^p, \quad [\text{Rao et al., 2003}]
\end{align*}
\]

♦ Different choices favor different levels of sparsity.
Example 2-D Contours

\[ g(x_i) = |x_i|^p \]
Which Penalty Should We Choose?

Two competing issues:

1. If the prior is too sparse (e.g., $p \approx 0$), then we may get stuck in a local minima: convergence error.

2. If the prior is not sparse enough (e.g. $p \approx 1$), then the global minimum may be found, but it might not equal $x_0$: structural error.

Answer is ultimately application- and algorithm-dependent
Convergence Errors vs. Structural Errors

Convergence Error \((p \approx 0)\)

Structural Error \((p \approx 1)\)

\(-\log p(x|y)\)

\(x'\) = solution we have converged to

\(x_0\) = maximally sparse solution
Desirable Properties of Algorithms

♦ Can be implemented via relatively simple primitives already in use, e.g., $\ell_1$ solvers, etc.

♦ Improved performance over OMP, BP, etc.

♦ Naturally extends to more general problems:

1. Constrained sparse estimation (e.g., finding non-negative sparse solutions)

2. Group sparsity problems …
Extension: Group Sparsity

Example:
- The *simultaneous sparse estimation problem* - the goal is to recover a matrix $X$, with maximal row sparsity [Cotter et al., 2005; Tropp, 2006], given observation matrix $Y$ produced via
  \[
  Y = \Phi X + E
  \]

Optimization Problem:

\[
X_0 (\lambda) = \arg \min_X \| Y - \Phi X \|_F^2 + \lambda \sum_{i=1}^m I[\| x_i \| \neq 0]
\]

- Can be efficiently solved/approximated by replacing indicator function with alternative function $g$. 

# of nonzero rows in $X$
Reweighted $\ell_2$ Optimization

- **Assume:** $g(x_i) = h(x_i^2)$, $h$ concave

- **Updates:**
  \[
  x^{(k+1)} \rightarrow \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_i w_i^{(k)} x_i^2 \\
  = \tilde{W}^{(k)} \Phi^T \left( \lambda I + \Phi \tilde{W}^{(k)} \Phi^T \right)^{-1} y
  \]

  \[
  w_i^{(k+1)} \rightarrow \frac{\partial g(x_i)}{\partial x_i^2} \bigg|_{x_i = x_i^{(k+1)}}, \quad \tilde{W}^{(k+1)} \rightarrow \text{diag} \left[ w^{(k+1)} \right]^{-1}
  \]

- Based on simple 1st order approx. to $g(x_i)$ [Palmer et al., 2006].
- Guaranteed not to increase objective function.
- Global convergence assured given additional assumptions.
Examples

1. **FOCUSS algorithm** [Rao et al., 2003]:
   - **Penalty:** \( g(x_i) = \left| x_i \right|^p, \ 0 \leq p \leq 2 \)
   - **Weight update:** \( w_i^{(k+1)} \rightarrow \left| x_i^{(k+1)} \right|^{p-2} \)
   - **Properties:** Well-characterized convergence rates; very susceptible to local minima when \( p \) is small.

2. **Chartrand and Yin (2008) algorithm**:
   - **Penalty:** \( g(x_i) = \log \left( x_i^2 + \varepsilon \right), \ \varepsilon \geq 0 \)
   - **Weight update:** \( w_i^{(k+1)} \rightarrow \left[ \left( x_i^{(k+1)} \right)^2 + \varepsilon \right]^{-1} \)
   - **Properties:** Slowly reducing \( \varepsilon \) to zero smoothes out local minima initially allowing better solutions to be found; very useful for recovering scaled coefficients ...
Empirical Comparison

For each test case:

1. Generate a random dictionary $\Phi$ with 50 rows and 250 columns.

2. Generate a sparse coefficient vector $x_0$.

3. Compute signal via $y = \Phi x_0$ (noiseless).

4. Compare Chartrand and Yin’s reweighted $\ell_2$ method with $\ell_1$ norm solution with regard to estimating $x_0$.

5. Average over 1000 independent trials.

Repeat with different sparsity levels and different nonzero coefficient distributions.
Empirical: Unit Nonzeros

![Graph showing the probability of success against $\|x_0\|_0$](image)
Results: Gaussian Nonzeros

- **Probability of Success**

Graph showing the relationship between the probability of success and $\|X_0\|_0$. The graph includes two curves:
- Red line: $L_1$ norm solution
- Blue line: Chartrand and Yin (2008)
Reweighted $\ell_1$ Optimization

- **Assume:**
  \[ g(x_i) = h(|x_i|), \ h \text{ concave} \]

- **Updates:**
  \[
  x^{(k+1)} \rightarrow \arg\min_{x} \|y - \Phi x\|_2^2 + \lambda \sum_i w_i^{(k)} |x_i| \\
  w_i^{(k+1)} \rightarrow \frac{\partial g(x_i)}{\partial |x_i|} \bigg|_{x_i = x_i^{(k+1)}}
  \]

- Based on simple 1st order approximation to $g(x_i)$ [Fazel et al., 2003]
- Global convergence given minimal assumptions [Zangwill, 1969].
- Per-iteration cost expensive, but few needed (and each are sparse).
- Easy to incorporate alternative data fit terms or constraints on $x$. 
Example [Candès et al., 2008]

- **Penalty:** \( g(x_i) = \log(|x_i| + \epsilon), \quad \epsilon \geq 0 \)

- **Updates:**
  \[
  x^{(k+1)} \rightarrow \arg \min_x \|y - \Phi x\|_2^2 + \lambda \sum_i w_i^{(k)} |x_i|
  \]
  \[
  w_i^{(k+1)} \rightarrow \left( |x_i^{(k+1)}| + \epsilon \right)^{-1}
  \]

- When nonzeros in \( x_0 \) are scaled, works much better than regular \( \ell_1 \), depending on how \( \epsilon \) is chosen.

- Local minima exist, but since each iteration is sparse, local solutions are not so bad (no worse than regular \( \ell_1 \) solution).
Empirical Comparison

- For each test case:
  1. Generate a random dictionary $\Phi$ with 50 rows and 100 columns.
  2. Generate a sparse coefficient vector $x_0$ with 30 truncated Gaussian, strictly positive nonzero coefficients.
  3. Compute signal via $y = \Phi x_0$ (noiseless).
  4. Compare Candes et al.’s reweighted $\ell_1$ method (10 iterations) with $\ell_1$ norm solution, both constrained to be non-negative to try and estimate $x_0$.
  5. Average over 1000 independent trials.

- Repeat with different values of the parameter $\varepsilon$. 
Empirical Comparison

The graph illustrates the probability of success as a function of $\epsilon$. The orange line represents Candes et al. (2008) and the blue line represents the $L_1$ norm solution. The x-axis represents $\epsilon$ ranging from $10^{-4}$ to $10^4$, while the y-axis shows the probability of success.
Conclusions

♦ In practice, MAP estimation addresses some limitations of standard methods (although not Limitation III, assessing uncertainty).

♦ Simple updates are possible using either iterative reweighted $l_1$ or $l_2$ minimization.

♦ More generally, iterative reweighted $f$ minimization, where $f$ is a convex function, is possible.
Section III: Bayesian Inference Using the Sparse Linear Model
Note

- MAP estimation is really just standard/classical penalized regression.

- So the Bayesian interpretation has not really contributed much as of yet ...
Posterior Modes vs. Posterior Mass

- Previous methods focus on finding the implicit *mode* of $p(x|y)$ by maximizing the joint distribution
  $$p(x, y) = p(y | x) p(x)$$

- *Bayesian inference* uses posterior information beyond the mode, i.e., posterior *mass*:

- Examples:
  1. *Posterior mean*: Can have attractive properties when used as a sparse point estimate (more on this later …).
  2. *Posterior covariance*: Useful assessing uncertainty in estimates, e.g., experimental design, learning new projection directions for compressive sensing measurements.
Posterior Modes vs. Posterior Mass

\[ p(x|y) \]

Mode

Probability Mass
Problem

♦ For essentially all sparse priors, cannot compute normalized posterior

\[ p(x \mid y) = \frac{p(y \mid x) p(x)}{\int p(y \mid x) p(x) \, dx} \]

♦ Also cannot compute posterior moments, e.g.,

\[
\begin{align*}
\mu_x &= E[x \mid y] \\
\Sigma_x &= \text{Cov}[x \mid y]
\end{align*}
\]

♦ So efficient approximations are needed …
Approximating Posterior Mass

- **Goal**: Approximate $p(x, y)$ with some distribution $\hat{p}(x, y)$ that
  
  1. Reflects the significant mass in $p(x, y)$.
  2. Can be normalized to get the posterior $p(x|y)$.
  3. Has easily computable moments, e.g., can compute $E[x|y]$ or $\text{Cov}[x|y]$.

- **Optimization Problem**: Find the $\hat{p}(x, y)$ that minimizes the sum of the misaligned mass:

\[
\int |p(x, y) - \hat{p}(x, y)| \, dx = \int p(y|x) |p(x) - \hat{p}(x)| \, dx
\]
Recipe

1. Start with a Gaussian likelihood

\[ p(y \mid x) = (2\pi \lambda)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \|y - \Phi x\|_2^2\right) \]

2. Pick an appropriate prior \( p(x) \) that encourages sparsity

3. Choose a convenient parameterized class of approximate priors \( \hat{p}(x) = p(x; \gamma) \)

4. Solve: \[ \hat{\gamma} = \arg\min_\gamma \int p(y \mid x) \left| p(x) - p(x; \gamma) \right| dx \]

5. Normalize: \[ p(x \mid y; \hat{\gamma}) = \frac{p(y \mid x) p(x; \hat{\gamma})}{\int p(y \mid x) p(x; \hat{\gamma}) dx} \]
Step 2: Prior Selection

- Assume a sparse prior distribution on each coefficient:

$$-\log p(x_i) \propto g(x_i) = h(x_i^2), \quad h \text{ concave, non-decreasing.}$$

2-D example

Gaussian Distribution

Sparse Distribution

[Tipping, 2001]
Step 3: Forming Approximate Priors $p(x; \gamma)$

- Any sparse prior can be expressed via the dual form [Palmer et al., 2006]:

$$ p(x_i) = \max_{\gamma_i \geq 0} \left[ (2\pi\gamma_i)^{-1/2} \exp\left(-\frac{x_i^2}{2\gamma_i}\right) \phi(\gamma_i) \right] $$

- **Two options**:
  1. Start with $p(x_i)$ and then compute $\phi(\gamma_i)$ via convexity results, or
  2. Choose $\phi(\gamma_i)$ directly and then compute $p(x_i)$; this procedure will always produce a sparse prior [Palmer et al. 2006].

- Dropping the maximization gives the strict variational lower bound

$$ p(x_i) \geq p(x_i; \gamma_i) = (2\pi\gamma_i)^{-1/2} \exp\left(-\frac{x_i^2}{2\gamma_i}\right) \phi(\gamma_i) $$

- Yields a convenient class of scaled Gaussian approximations:

$$ p(x; \gamma) = \prod_i p(x_i; \gamma_i) $$
Example: Approximations to Sparse Prior

\[ p(x_i) \]

\[ p(x_i; \gamma_i) \]
Step 4: Solving for the Optimal $\gamma$

- To find the best approximation, must solve

$$\hat{\gamma} = \arg\min_{\gamma \geq 0} \int p(y \mid x) \left| p(x) - p(x; \gamma) \right| dx$$

- By virtue of the strict lower bound, this is equivalent to

$$\hat{\gamma} = \arg\max_{\gamma \geq 0} \int p(y \mid x) p(x; \gamma) dx$$

$$= \arg\min_{\gamma \geq 0} \log \left| \Sigma_y \right| + y^T \Sigma_y^{-1} y - 2 \sum_{i=1}^{m} \log \varphi(\gamma_i)$$

where $\Sigma_y = \lambda I + \Phi \Gamma \Phi^T$ \hspace{1cm} $\Gamma = \text{diag}(\gamma)$
How difficult is finding the optimal parameters $\gamma$?

- If original MAP estimation problem is convex, then so is Bayesian inference cost function [Nickisch and Seeger, 2009].

- In other situations, Bayesian inference cost is generally much smoother than associated MAP estimation (more on this later …).
Step 5: Posterior Approximation

♦ We have found the approximation

\[ p(y \mid x) p(x; \hat{\gamma}) = p(x, y; \hat{\gamma}) \approx p(x, y) \]

♦ By design, this approximation reflects significant mass in the full distribution \( p(x,y) \).

♦ Also, it is easily normalized to form

\[ p(x \mid y; \hat{\gamma}) = N(\mu_x, \Sigma_x) \]

\[
\begin{align*}
\mu_x & = \mathbb{E}[x \mid y; \hat{\gamma}] = \hat{\Gamma} \Phi^T \left( \lambda I + \Phi \hat{\Gamma} \Phi^T \right)^{-1} y \\
\Sigma_x & = \text{Cov}[x \mid y; \hat{\gamma}] = \hat{\Gamma} - \hat{\Gamma} \Phi^T \left( \lambda I + \Phi \hat{\Gamma} \Phi^T \right)^{-1} \Phi \hat{\Gamma}
\end{align*}
\]
Applications of Bayesian Inference

1. Finding maximally sparse representations
   - Replace MAP estimate with posterior mean estimator.
   - For certain prior selections, this can be very effective (next section)

2. Active learning, experimental design
Experimental Design

♦ **Basic Idea** [Shihao Ji et al., 2007, Seeger and Nickisch, 2008]:
Use the approximate posterior

\[ p(\mathbf{x} | y; \hat{\gamma}) = N(\mu_x, \Sigma_x) \]

to learn new rows of the design matrix \( \Phi \) such that uncertainty about \( \mathbf{x} \) is reduced.

♦ Choose each additional row to minimize the differential entropy \( H \):

\[
H = \frac{1}{2} \log |\Sigma_x|, \quad \Sigma_x = \hat{\Gamma} - \hat{\Gamma} \Phi^T \left( \lambda I + \Phi \hat{\Gamma} \Phi^T \right)^{-1} \Phi \hat{\Gamma}
\]
Experimental Design Cont.

- Drastic improvement over random projections is possible in a variety of domains.

- Examples:
  - Reconstructing natural scenes [Seeger and Nickisch, 2008]
  - Undersampled MRI reconstruction [Seeger et al., 2009]
Section IV: Analysis of Bayesian Inference and Connections with MAP
Overview

- Bayesian inference can be recast as a general form of MAP estimation in \( x \)-space.

- This is useful for several reasons:
  1. Allows us to leverage same algorithmic formulations as with iterative reweighted methods for MAP estimation.
  2. Reveals that Bayesian inference can actually be an easier computational task than searching for the mode as with MAP.
  3. Provides theoretical motivation for posterior mean estimator when searching for maximally sparse solutions.
  4. Allows modifications to Bayesian inference cost (e.g., adding constraints), and inspires new non-Bayesian sparse estimators.
Reformulation of Posterior Mean Estimator

Theorem

\[
\mu_x = E[x \mid y, \hat{\gamma}] = \arg \min_x \| y - \Phi x \|_2^2 + \lambda g_{\text{infer}}(x)
\]

with Bayesian inference penalty function

\[
g_{\text{infer}}(x) = \min_{\gamma \geq 0} x^T \Gamma^{-1} x + \log |\lambda I + \Phi \Gamma \Phi^T| - 2 \sum_i \log \phi(\gamma_i)
\]

[Wipf and Nagarajan, 2008]

- So the posterior mean can be obtained by minimizing a penalized regression problem just like MAP
- Posterior covariance is a natural byproduct as well
Property I of Penalty $g_{\text{infer}}(x)$

- Penalty $g_{\text{infer}}(x)$ is formed from a minimum of upper-bounding hyperplanes with respect to each $x_i^2$.

This implies:

1. Concavity in $x_i^2$ for all $i$ [Boyd 2004].

2. Can be implemented via iterative reweighted $\ell_2$ minimization (multiple possibilities using various bounding techniques) [Seeger, 2009; Wipf and Nagarajan, 2009].

3. Note: Posterior covariance can be obtained easily too, therefore entire approximation can be computed for full Bayesian inference.
Student’s t Example

- Assume the following sparse distribution for each unknown coefficient:

\[ p(x_i) \propto \left( b + \frac{x_i^2}{2} \right)^{-(a+1/2)} \]

Note:
- When \( a = b \to \infty \), prior approaches a Gaussian (not sparse)
- When \( a = b \to 0 \), prior approaches a Jeffreys (highly sparse)

- Using convex bounding techniques to approximate the required derivatives, leads to simple reweighted \( \ell_2 \) update rules [Seeger, 2009; Wipf and Nagarajan, 2009].

- Algorithm can also be derived via EM [Tipping, 2001].
Reweighted $\ell_2$ Implementation Example

$$\mathbf{x}^{(k+1)} \rightarrow \arg \min_{\mathbf{x}} \| \mathbf{y} - \Phi \mathbf{x} \|^2_2 + \lambda \sum_i w_i^{(k)} x_i^2$$

$$= \tilde{\mathbf{W}}^{(k)} \Phi^T (\lambda I + \Phi \tilde{\mathbf{W}}^{(k)} \Phi^T)^{-1} \mathbf{y}, \quad \tilde{\mathbf{W}}^{(k)} = \text{diag}[\mathbf{w}^{(k)}]^{-1}$$

$$w_i^{(k+1)} \rightarrow \frac{1+2a}{\left( x_i^{(k+1)} \right)^2 + \left( w_i^{(k)} \right)^{-1} - \left( w_i^{(k)} \right)^{-2} \phi_i^T (\lambda I + \Phi \tilde{\mathbf{W}}^{(k)} \Phi^T)^{-1} \phi_i + 2b}$$

- Guaranteed to reduce or leave unchanged objective function at each iteration
- Other variants are possible using different bounding techniques
- Upon convergence, posterior covariance is given by

$$\Sigma_x = \left[ \lambda^{-1} \Phi^T \Phi + W^{(k)} \right]^{-1}, \quad W^{(k)} = \text{diag}[\mathbf{w}^{(k)}]$$
Property II of Penalty $g_{\text{infer}}(x)$

If $-2 \log \varphi(\gamma_i)$ is concave in $\gamma_i$, then:

1. $g_{\text{infer}}(x)$ is concave in $|x_i|$ for all $i$  \quad \Rightarrow \quad \text{sparsity-inducing}$

2. The implies posterior mean will always have at least $m - n$ elements equal to exactly zero as occurs with MAP.

3. Can be useful for canonical sparse recovery problems …

4. Can implement via reweighted $\ell_1$ minimization

[Wipf, 2006; Wipf and Nagarajan, 2009]
Reweighted $\ell_1$ Implementation Example

♦ Assume $-2\log \varphi(\gamma_i) = a\gamma_i$, $a \geq 0$

\[
x^{(k+1)} \rightarrow \arg\min_x \|y - \Phi x\|_2^2 + \lambda \sum_i w_i^{(k)} |x_i|
\]

\[
w_i^{(k+1)} \rightarrow \left[ \phi_i^T \left( \sigma^2 I + \Phi \tilde{W}^{(k)} \text{diag}\left[ x^{(k+1)} \right] \Phi^T \right)^{-1} \phi_i + a \right]^{1/2}
\]

- Guaranteed to converge to a minimum of the Bayesian inference objective function
- Easy to outfit with additional constraints
- In noiseless case, sparsity will not increase, i.e.,

\[
\|x^{(k+1)}\|_0 \leq \|x^{(k)}\|_0 \leq \|x^{(\ell_1\text{-norm})}\|_0
\]
Property III of Penalty $g_{\text{infer}}(x)$

- Bayesian inference is most sensitive to posterior mass, therefore it is less sensitive to spurious local peaks as is MAP estimation.

- Consequently, in $x$-space, the Bayesian Inference penalized regression problem

$$\min_x \|y - \Phi x\|_2^2 + \lambda g_{\text{infer}}(x)$$

is generally smoother than associated MAP problem.
Student’s t Example

♦ Assume the Student’s t distribution for each unknown coefficient:

\[ p(x_i) \propto \left( b + \frac{x_i^2}{2} \right)^{-(a+1/2)} \]

Note:

♦ When \( a = b \to \infty \), prior approaches a Gaussian (not sparse)
♦ When \( a = b \to 0 \), prior approaches a Jeffreys (highly sparse)

♦ Goal: Compare Bayesian inference and MAP for different sparsity levels to show smoothing effect.
Visualizing Local Minima Smoothing

- Consider when \( y = \Phi x \) has a 1-D feasible region, i.e.,
  \[ m = n + 1 \]

- Any feasible solution \( x \) will satisfy:
  \[ x = x_{\text{true}} + \alpha v \]
  \( v \in \text{Null}(\Phi) \)
  where \( \alpha \) is a scalar
  \( x_{\text{true}} \) is the true generative coefficients

- Can plot penalty functions vs. \( \alpha \) to view local minima profile over the 1-D feasible region.
Empirical Example

- Generate an iid Gaussian random dictionary $\Phi$ with 10 rows and 11 columns.

- Generate a sparse coefficient vector $x_{\text{true}}$ with 9 nonzeros and Gaussian iid amplitudes.

- Compute signal $y = \Phi x_0$.

- Assume a Student’s t prior on $x$ with varying degrees of sparsity.

- Plot MAP/Bayesian inference penalties vs. $\alpha$ to compare local minima profiles over the 1-D feasible region.
Local Minima Smoothing Example #1

Low Sparsity
Student’s t

penalty, normalized

\( \alpha \)
Local Minima Smoothing Example #2

Medium Sparsity
Student’s $t$

penalty, normalized

$\alpha$

Inference
MAP
Local Minima Smoothing Example #3

High Sparsity
Student’s t

penalty, normalized vs. $\alpha$

Inference
MAP
Property IV of Penalty $g_{\text{infer}}(x)$

- **Non-separable**, meaning $g_{\text{infer}}(x) \neq \sum_i g_i(x_i)$

- Non-separable penalty functions can have an advantage over separable penalties (i.e., MAP) when it comes to canonical sparse recovery problems [Wipf and Nagarajan, 2010].
Example

- Consider original sparse estimation problem

\[ x_0 \triangleq \arg \min_x \|x\|_0 \quad \text{s.t.} \quad y = \Phi x \]

- **Problem**: Combinatorial number of local minima:

\[
\binom{m-1}{n} + 1 \leq \text{number of local minima} \leq \binom{m}{n}
\]

- Local minima occur at each basic feasible solution (BFS):

\[ \|x\|_0 \leq n \quad \text{s.t.} \quad y = \Phi x \]
Visualization of Local Minima in $\ell_0$ Norm

- Generate an iid Gaussian random dictionary $\Phi$ with 10 rows and 11 columns.

- Generate a sparse coefficient vector $x_{\text{true}}$ with 9 nonzeros and Gaussian iid amplitudes.

- Compute signal via $y = \Phi x_0$.

- Plot $\|x\|_0$ vs. $\alpha$ (1-D null space dimension) to view local minima profile of the $\ell_0$ norm over the 1-D feasible region.
$\ell_0$ Norm in 1-D Feasible Region
Non-Separable Penalty Example

♦ Would like to smooth local minima while retaining same global solution as $\ell_0$ at all times (unlike $\ell_1$ norm)

♦ This can be accomplished by a simple modification of the $\ell_0$ penalty.

♦ **Truncated $\ell_0$ penalty:**

$$\hat{x} = \arg \min_x \|\tilde{x}\|_0 \quad \text{s.t.} \quad \tilde{x} = k \text{ largest elements of } x$$

$$y = \Phi x$$

♦ If $k < m$, then there will necessarily be fewer local minima; however, the implicit prior/penalty function is *non-separable*. 
Truncated $\ell_0$ Norm in 1-D Feasible Region

# of nonzeros

$\alpha$

$\text{regular } L_0 \text{ norm}$

$\text{truncated } L_0 \text{ norm}$

$k=10$
Using posterior mean estimator for finding maximally sparse solutions

Summary of why this could be a good idea:

1. If \(-2 \log \phi(\gamma_i)\) is concave, then posterior mean will be sparse (local minima of Bayesian inference cost will also be sparse).

2. The implicit Bayesian inference cost function can be much smoother than the associated MAP objective.

3. Potential advantages of non-separable penalty functions.
Choosing the function $\varphi$

- For sparsity, require that $-2\log \varphi(\gamma_i)$ is concave.

- To avoid adding extra local minima (i.e., to maximally exploit smoothing effect), require that $-2\log \varphi(\gamma_i)$ is convex.

- So $-2\log \varphi(\gamma_i) = a\gamma_i$, $a \geq 0$ is well-motivated [Wipf et al. 2007].

Assume simplest case: $-2\log \varphi(\gamma_i) = 0$, sometimes referred to as sparse Bayesian learning (SBL) [Tipping, 2001].

- We denote the penalty function in this case $g_{\text{SBL}}(x)$. 
Advantages of Posterior Mean Estimator

**Theorem**

- In the low noise limit \((\lambda \rightarrow 0)\), and assuming \(\|x_0\|_0 < \text{spark}[\Phi] - 1\), then the SBL penalty satisfies:

  \[
  \arg\min_{x: y = \Phi x} g_{\text{SBL}}(x) = \arg\min_{x: y = \Phi x} \|x\|_0
  \]

- No separable penalty \(g(x) = \sum_i g_i(x_i)\) satisfies this condition and has fewer minima than the SBL penalty in the feasible region.

[Wipf and Nagarajan, 2008]
Conditions For a Single Minimum

Theorem

- Assume $\|x_0\|_0 < \text{spark}[\Phi] - 1$. If the magnitudes of the non-zero elements in $x_0$ are sufficiently scaled, then the SBL cost ($\lambda \to 0$) has a \textit{single minimum} which is located at $x_0$.

No possible separable penalty (standard MAP) satisfies this condition.

[Wipf and Nagarajan, 2009]
Empirical Example

- Generate an iid Gaussian random dictionary $\Phi$ with $10$ rows and $11$ columns.

- Generate a maximally sparse coefficient vector $x_0$ with $9$ nonzeros and either
  1. amplitudes of similar scales, or
  2. amplitudes with very different scales.

- Compute signal via $y = \Phi x_0$.

- Plot MAP/Bayesian inference penalty functions vs. $\alpha$ to compare local minima profiles over the 1-D feasible region to see the effect of coefficient scaling.
Smoothing Example: Similar Scales
Smoothing Example: Different Scales

penalty, normalized

\( \alpha \)
Consider the noiseless sparse recovery problem.
\[ x_0 \triangleq \arg \min_x \|x\|_0 \text{ s.t. } y = \Phi x \]

Under very mild conditions, SBL with reweighted $\ell_1$ implementation will:

1. Never do worse than the regular $\ell_1$-norm solution
2. For any dictionary and sparsity profile, there will always be cases where it does better.

[Wipf and Nagarajan, 2010]
Empirical Example: 
Simultaneous Sparse Approximation

♦ Generate data matrix via $Y = \Phi X_0$ (noiseless):

♦ $X_0$ is 100-by-5 with random nonzero rows.

♦ $\Phi$ is 50-by-100 with Gaussian iid entries

♦ Check if $X_0$ is recovered using various algorithms:

1. Generalized SBL, reweighted $\ell_2$ implementation [Wipf and Nagarajan, 2010]
2. Candès et al. (2008) reweighted $\ell_1$ method
3. Chartrand and Yin (2008) reweighted $\ell_2$ method
4. $\ell_1$ solution via Group Lasso [Yuan and Lin, 2006]
Conclusions

♦ Posterior information beyond the mode can be very useful in a wide variety of applications.

♦ Variational approximation provides useful estimates of posterior means and covariances, which can be computed efficiently using standard iterative reweighting algorithms.

♦ In certain situations, posterior mean estimate can be effective substitute for $\ell_0$ norm minimization.

♦ In simulation tests, out-performs a wide variety of MAP-based algorithms [Wipf and Nagarajan, 2010]...
Section V: Application Examples in Neuroimaging
Applications of Sparse Bayesian Methods

1. Recovering fiber track geometry from diffusion weighted MR images [Ramirez-Manzanares et al. 2007].


3. Compressive sensing for rapid MRI [Lustig et al. 2007].

4. MEG/EEG source localization [Sato et al. 2004; Friston et al. 2008].
MEG/EEG Source Localization

Maxwell’s eqs.

source space (X) → sensor space (Y)

?
The Dictionary $\Phi$

- Can be computed using a boundary element brain model and Maxwell’s equations.

- Will be dependent on location of sensors and whether we are doing MEG, EEG, or both.

- Unlike compressive sensing domain, columns of $\Phi$ will be highly correlated regardless of where sensors are placed.
Source Localization

♦ Given multiple measurement vectors $Y$, MAP or Bayesian inference algorithms can be run to estimate $X$.

♦ The estimated nonzero rows should correspond with the location of active brain areas (also called sources).

♦ Like compressive sensing, may apply algorithms in appropriate transform domain where row-sparsity assumption holds.
Empirical Results

1. Simulations with real brain noise/interference:
   ♦ Generate damped sinusoidal sources
   ♦ Map to sensors using $\Phi$ and apply real brain noise, artifacts

2. Data from real-world experiments:
   ♦ Auditory evoked fields from binaurally presented tones (which produce correlated, bilateral activations)

Compare localization results using MAP estimation and SBL posterior mean from Bayesian inference
MEG Source Reconstruction Example

Ground Truth  SBL  Group Lasso
Real Data:
Auditory Evoked Field (AEF)

SBL

Beamformer

sLORETA

Group Lasso
Conclusion

- MEG/EEG source localization demonstrates the effectiveness of Bayesian inference on problems where the dictionary is:
  - Highly overcomplete, meaning $m \gg n$, e.g.,
    $$n = 275 \quad \text{and} \quad m = 100,000.$$
  - Very ill-conditioned and coherent, i.e., columns are highly correlated.
Final Thoughts

♦ *Sparse Signal Recovery* is an interesting area with many potential applications.

♦ Methods developed for solving the Sparse Signal Recovery problem can be valuable tools for signal processing practitioners.

♦ Rich set of computational algorithms, e.g.,
  ♦ Greedy search (OMP)
  ♦ $\ell_1$ norm minimization (Basis Pursuit, Lasso)
  ♦ MAP methods (Reweighted $\ell_1$ and $\ell_2$ methods)
  ♦ Bayesian Inference methods like SBL (show great promise)

♦ Potential for great theory in support of performance guarantees for algorithms.

♦ Expectation is that there will be continued growth in the application domain as well as in the algorithm development.
Thank You