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
  - ℓ_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
  - ℓ_1 norm minimization
- Performance Guarantees

Early Works

- R. R. Hocking and R. N. Leslie , "Selection of the Best Subset in Regression Analysis," Technometrics, 1967.
- S. Singhal and B. S. Atal, "Amplitude Optimization and Pitch Estimation in Multipulse Coders," *IEEE Trans. Acoust., Speech, Signal Processing*, 1989
- S. D. Cabrera and T. W. Parks, "Extrapolation and spectral estimation with iterative weighted norm modification," *IEEE Trans. Acoust., Speech, Signal Processing, April 1991.*
- Many More works
- Our first work
  - I.F. Gorodnitsky, B. D. Rao and J. George, "Source Localization in Magnetoencephalography using an Iterative Weighted Minimum Norm Algorithm, IEEE Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Pages: 167-171, Oct. 1992

Early Session on Sparsity

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

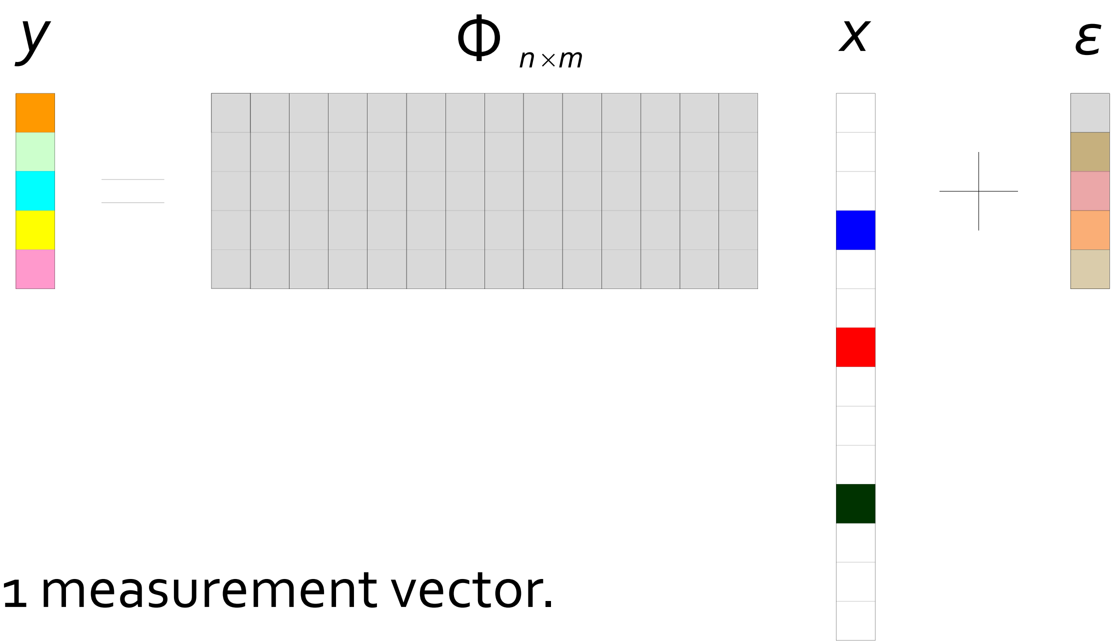
Motivation for Tutorial

- 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
  - ℓ_1 norm minimization
- Performance Guarantees

Problem Description



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

Problem Statement

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

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

where *I*(.) is the indicator function

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

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

Complexity

- Search over all possible subsets, which would mean a search over a total of $({}^{m}C_{k})$ subsets. Combinatorial Complexity.

  With $m = 30; n = 20$; and $k = 10$ there are 3×10^7 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
  - ℓ_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***

$$y[l] = 2(\cos\omega_0 l + \cos\omega_1 l),\ l = 0,1,2,\dots,n-1.\ n = 64.$$

$$\omega_0 = \frac{2\pi}{64}\frac{33}{2},\quad \omega_1 = \frac{2\pi}{64}\frac{34}{2}.$$

- **Dictionary Elements**:

$$\phi_l^{(m)} = [1, e^{-j\omega_l}, e^{-j2\omega_l}, \dots, e^{-j(n-1)\omega_l}]^T, \omega_l = \frac{2\pi}{m} l$$

- 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

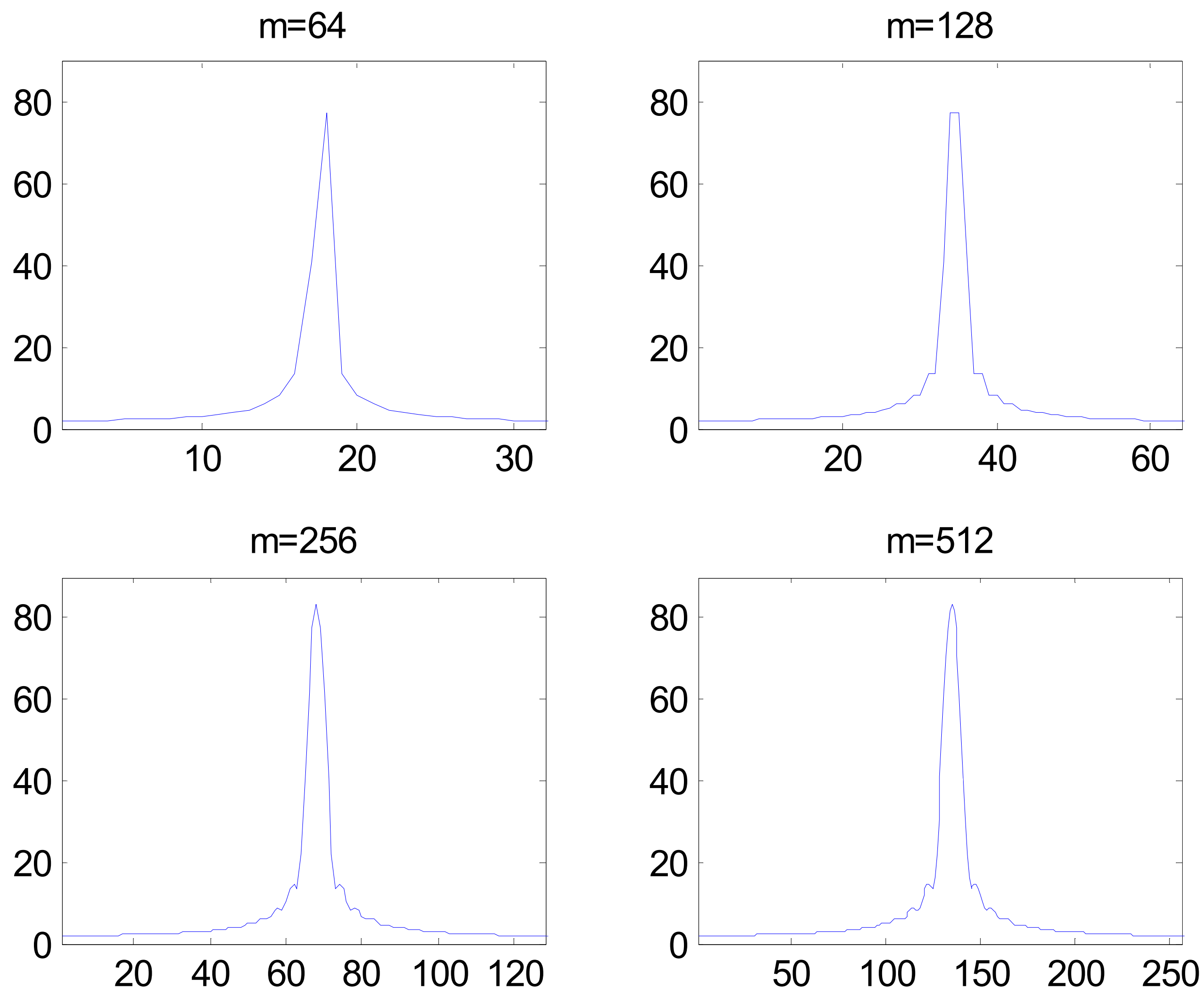
$$\begin{aligned} y &= \phi_{33}^{(128)} + \phi_{34}^{(128)} + \phi_{94}^{(128)} + \phi_{95}^{(128)} \\ &= \phi_{66}^{(256)} + \phi_{68}^{(256)} + \phi_{188}^{(256)} + \phi_{190}^{(256)} \\ &= \phi_{132}^{(512)} + \phi_{136}^{(512)} + \phi_{376}^{(512)} + \phi_{380}^{(512)} \end{aligned}$$

- Consider the linear system of equations

$$y = \Phi^{(m)} x$$

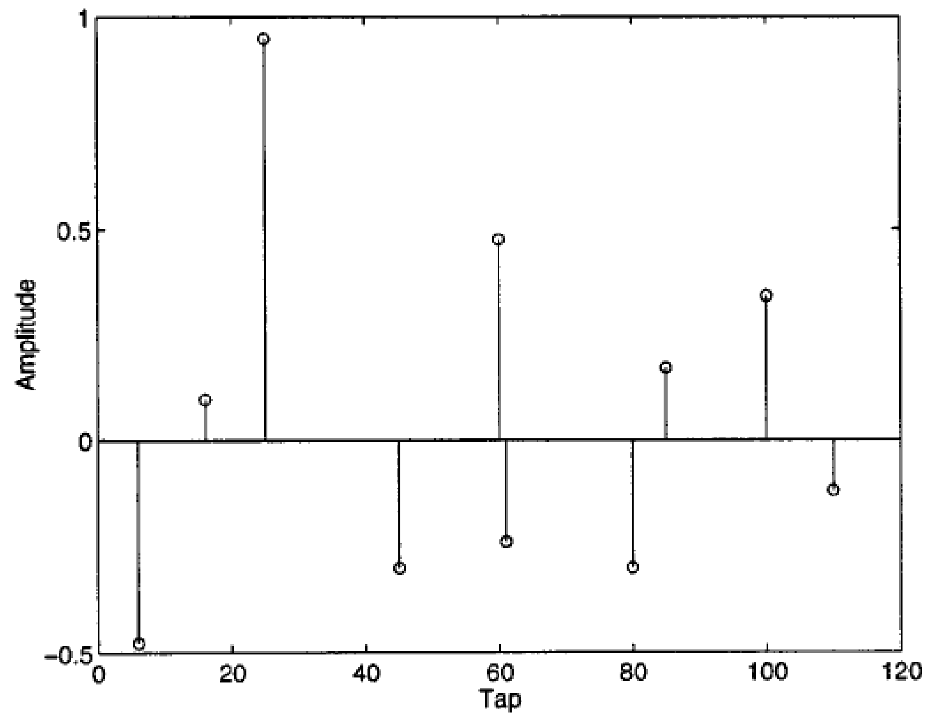
- The frequency components in the data are in the dictionaries $\Phi^{(m)}$ for $m = 128, 256, 512$.
- What solution among all possible solutions does the DFT compute?

DFT Example



m=64
m=128
m=256
m=512
80
60
40
20
0
10
20
30
40
60
50
100
120
150
200
250

Sparse Channel Estimation



Noise

$$r(i) = \sum_{j=0}^{m-1} s(i-j)c(j) - \varepsilon(i), \quad i = 0,1,\ldots,n-1$$

Received seq. Training seq. Channel impulse response

Example:
Sparse Channel Estimation

- Formulated as a sparse signal recovery problem

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

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

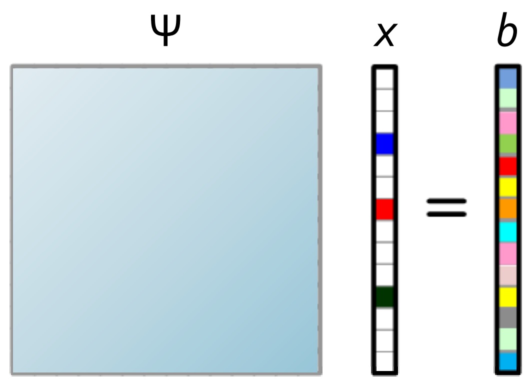
Compressive Sampling

- D. Donoho, “Compressed Sensing,” IEEE Trans. on Information Theory, 2006

- E. Candes and T. Tao, “Near Optimal Signal Recovery from random Projections: Universal Encoding Strategies,” IEEE Trans. on Information Theory, 2006

Compressive Sampling

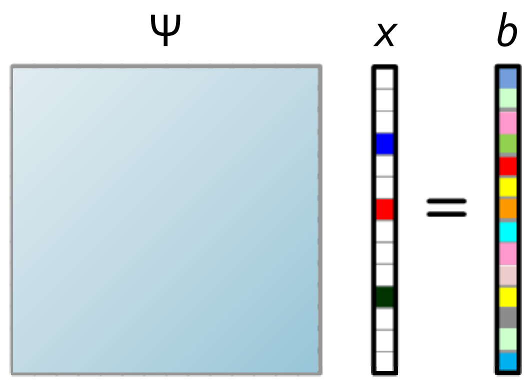
- Transform Coding



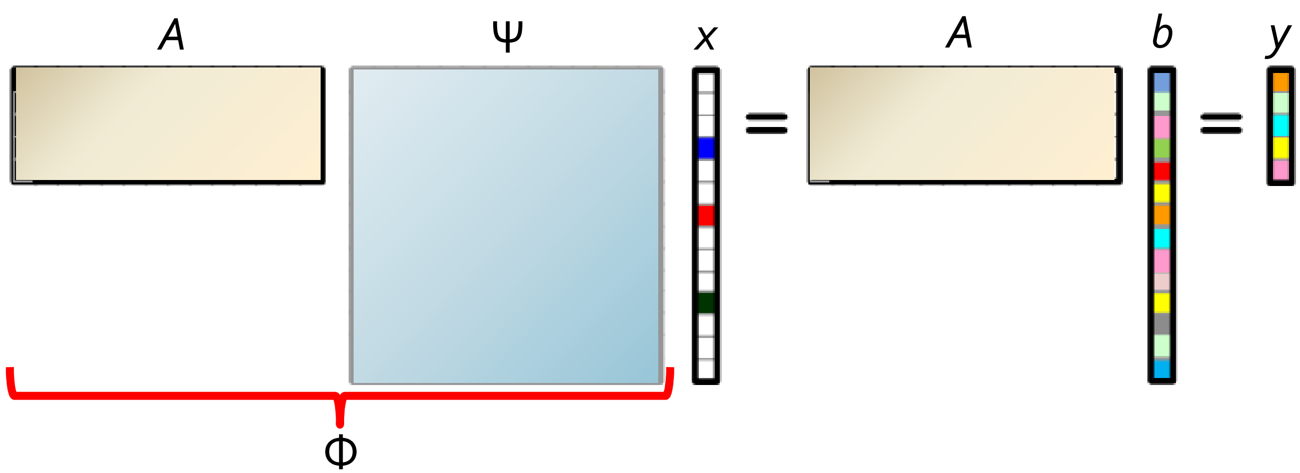
- What is the problem here?
  - Sampling at the Nyquist rate
  - Keeping only a small amount of nonzero coefficients
  - Can we directly acquire the signal below the Nyquist rate?

Compressive Sampling

- Transform Coding

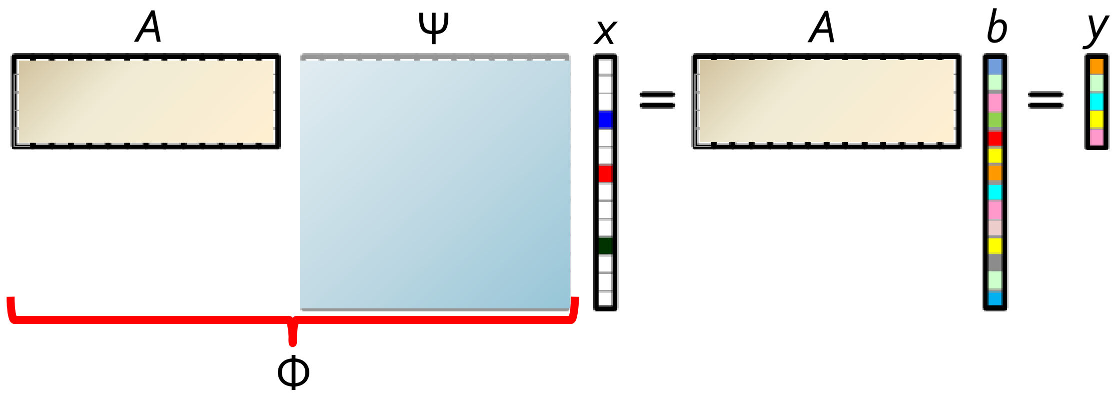


- Compressive Sampling



Compressive Sampling

- Compressive Sampling



- Computation:
  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
  - ℓ_1 norm minimization
- Performance Guarantees

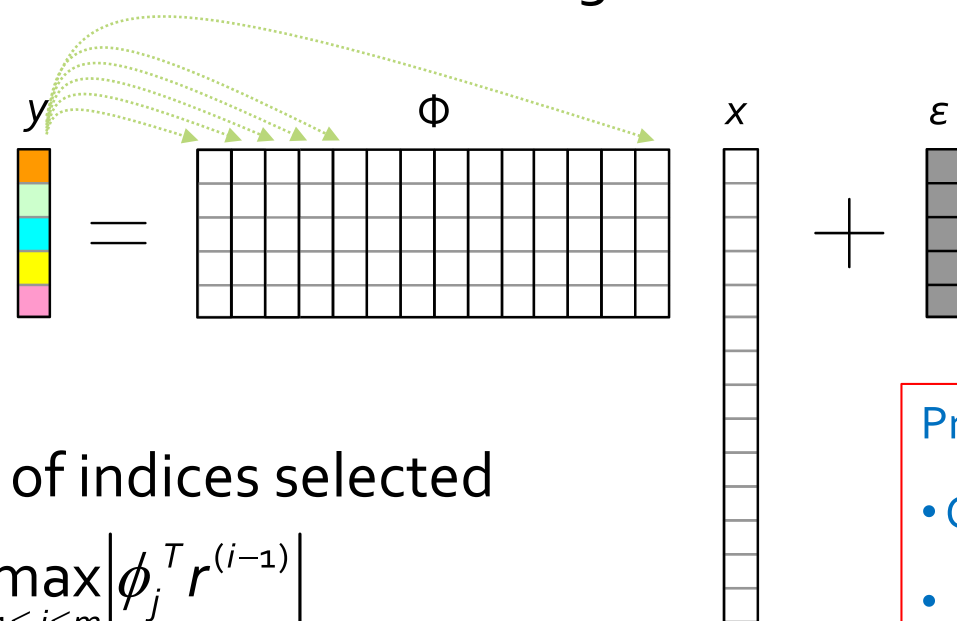
Potential Approaches

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. ℓ_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



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

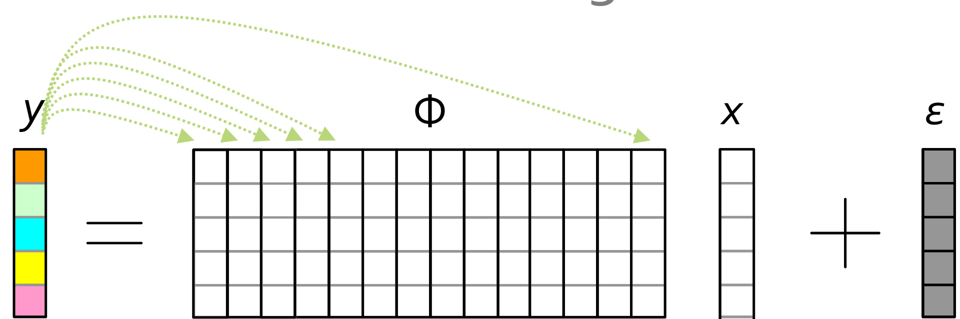
Practical stop criteria:

- Certain # iterations
- $\left\| r^{(i)} \right\|_2$ smaller than threshold

- Remove its contribution from the residual
  - Update $S^{(i)}$: If $l \notin S^{(i-1)}, S^{(i)} = S^{(i-1)} \bigcup \{l\}$. Or, keep $S^{(i)}$ the same
  - Update $r^{(i)}$: $r^{(i)} = \mathrm{P}_{\phi_l}^{\perp} r^{(i-1)} = r^{(i-1)} - \phi_l \phi_l^T r^{(i-1)}$

Greedy Search Method: Orthogonal Matching Pursuit (OMP)

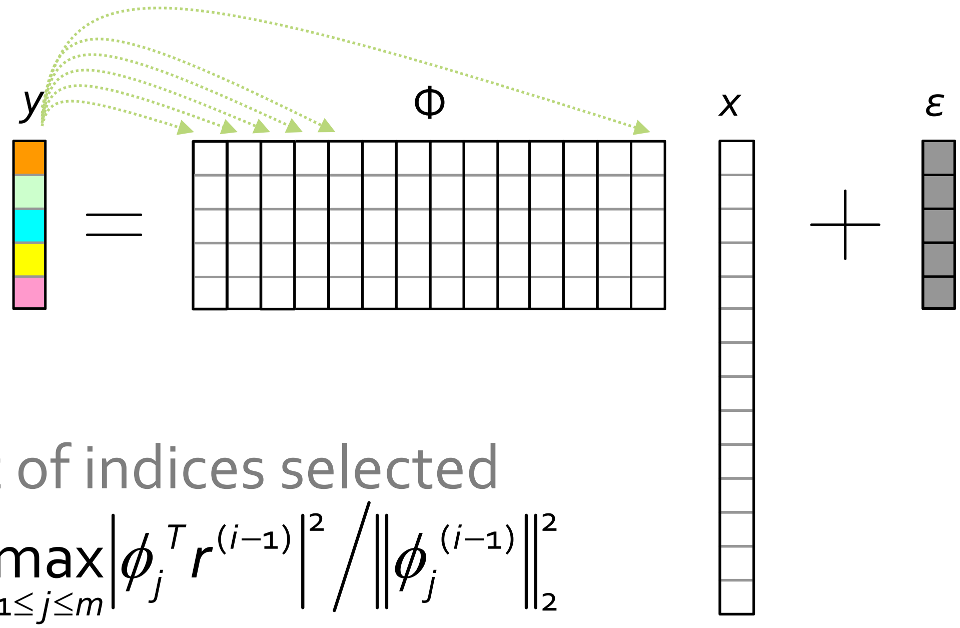
- Select a column that is most aligned with the current residual



  - $r^{(0)} = y$
  - $S^{(i)}$: set of indices selected
  - $l = \underset{1 \le j \le m}{\operatorname{argmax}} \left| \phi_j^T r^{(i-1)} \right|$
- Remove its contribution from the residual
  - Update $S^{(i)}$: $S^{(i)} = S^{(i-1)} \bigcup \{l\}$
  - Update $r^{(i)}$: $r^{(i)} = \mathrm{P}^{\perp}_{[l_1, l_2, \ldots, l_i]} r^{(i-1)} = r^{(i-1)} - \mathrm{P}_{[l_1, l_2, \ldots, l_i]} r^{(i-1)}$

Greedy Search Method: Order Recursive OMP

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



  - $r^{(0)} = y$
  - $S^{(i)}$: set of indices selected
  - $l = \underset{1 \le j \le m}{\operatorname{argmax}} \left| \phi_j^T r^{(i-1)} \right|^2 \Big/ \left\| \phi_j^{(i-1)} \right\|_2^2$
- Remove its contribution from the residual
  - Update $S^{(i)}$: $S^{(i)} = S^{(i-1)} \bigcup \{l\}$
  - Update $r^{(i)}$: $r^{(i)} = \mathrm{P}^{\perp}_{[l_1, l_2, \ldots, l_i]} r^{(i-1)} = r^{(i-1)} - \mathrm{P}_{[l_1, l_2, \ldots, l_i]} r^{(i-1)}$
  - Update $\left\| \phi_l^{(i)} \right\|_2^2$: $\phi_l^{(i)} = \mathrm{P}^{\perp}_{[l_1, l_2, \ldots, l_i]} \phi_l$. 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 Φ and $\Phi^+ = \Phi^T(\Phi\Phi^T)^{-1}$.

- Minimum Norm solution: The minimum ℓ_2 norm solution

  $x_{mn} = \Phi^+ y$ is a popular solution

- Noisy Case: regularized ℓ_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 ℓ_2 norm solution is not sparse

Example:

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

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

DFT: Also computes minimum 2-norm solution

Diversity Measures

- Recall:

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

- 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 $p \rightarrow 0$,

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

- ℓ_1 norm, convex relaxation of ℓ_0

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

ℓ_1 Diversity Measure

- Noiseless case

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

- Noisy case
  - ℓ_1 regularization [Candes, Romberg, Tao]

$$\min_x \sum_{i=1}^{m} |x_i| \ \text{ subject to } \ \|y - \Phi x\|_2 \le \beta$$

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

$$\min_x \|y - \Phi x\|_2^2 + \lambda \sum_{i=1}^{m} |x_i|$$

ℓ_1 norm minimization and MAP estimation

- MAP estimation

$$\hat{x} = \underset{x}{\operatorname{argmax}}\, p(x \mid y)$$
$$= \underset{x}{\operatorname{argmax}}\left[\log p(y \mid x) + \log p(x)\right]$$

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

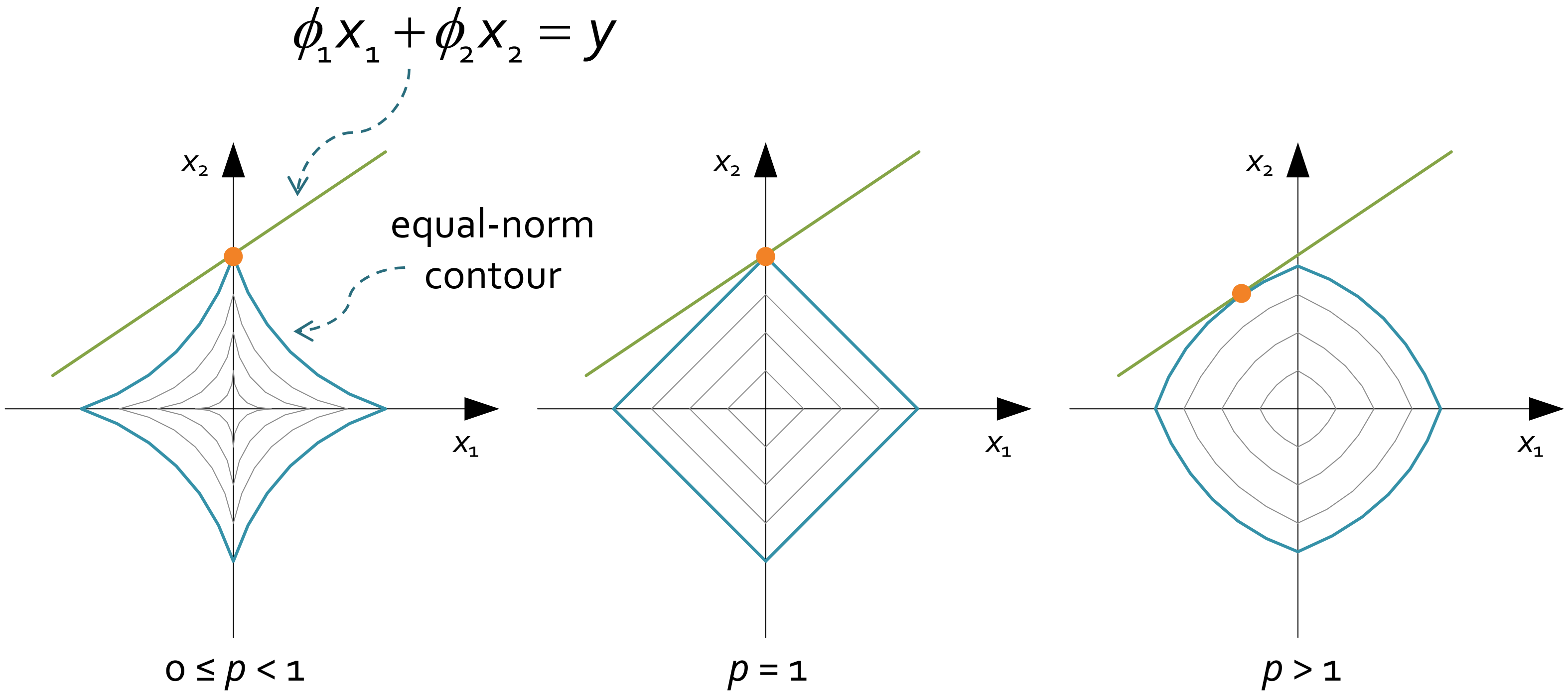
$$\hat{x} = \underset{x}{\operatorname{argmin}}\left[\left\|y - \Phi x\right\|_2^2 + \lambda \sum_i \left|x_i\right|\right]$$

Attractiveness of ℓ_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?

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



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 Φ is n×m, then a basic feasible solution is a solution with n non-zero entries

Example with ℓ_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 $\varepsilon = [0.01, \ -0.01]^T$
  - ℓ_1 regularization result: $x_{l_1R} = [0.986, \ 0, \ 8.77 \times 10^{-6}]^T$
  - Lasso result ($\lambda = 0.05$): $x_{lasso} = [0.975, \ 0, \ 2.50 \times 10^{-5}]^T$

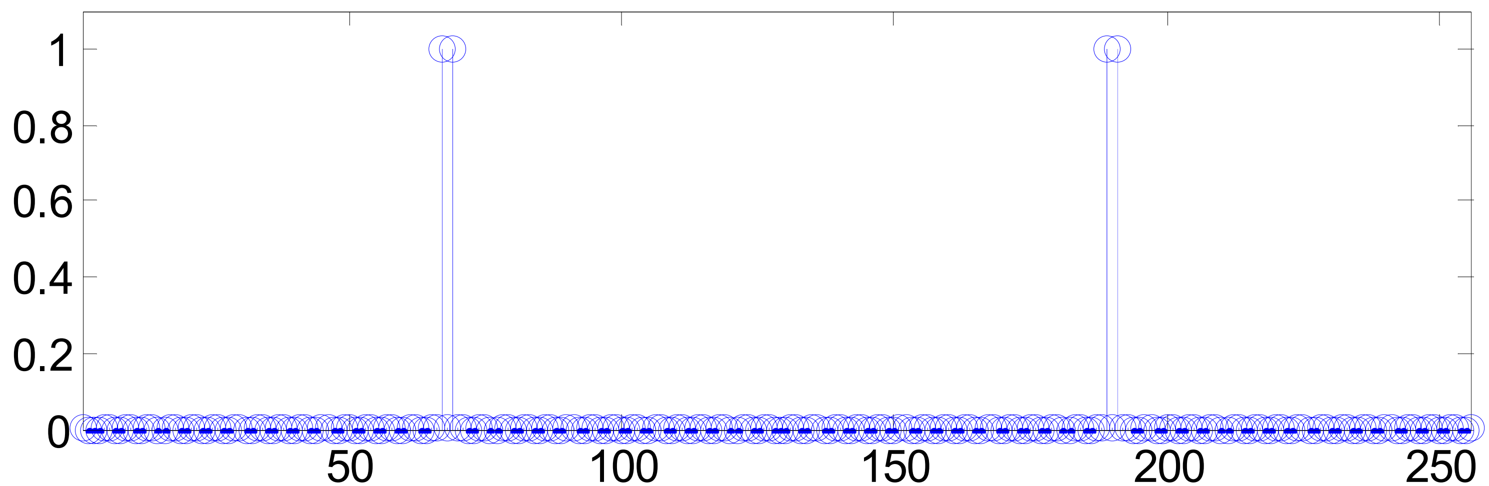
Example with ℓ_1 diversity measure

- Continue with the DFT example:

$$y[l] = 2(\cos\omega_0 l + \cos\omega_1 l),\ l = 0,1,2,\ldots,n-1.\ n = 64.$$

$$\omega_0 = \frac{2\pi}{64}\frac{33}{2},\quad \omega_1 = \frac{2\pi}{64}\frac{34}{2}.$$

- 64, 128,256,512 DFT cannot separate the adjacent frequency components
- Using ℓ_1 diversity measure minimization (m=256)



Outline: Part 1

- Motivation for Tutorial
- Sparse Signal Recovery Problem
- Applications
- Computational Algorithms
  - Greedy Search
  - ℓ_1 norm minimization
- Performance Guarantees

Important Questions

- When is the ℓ_0 solution unique?
- When is the ℓ_1 solution equivalent to that of ℓ_0?
  - Noiseless Case
  - Noisy Measurements

Uniqueness

- Definition of Spark
  - The smallest number of columns from Φ that are linearly dependent. (Spark(Φ) $\leq$ n+1)
- Uniqueness of sparsest solution
  - If $\sum_{i=1}^{m} I(x_i \neq 0) < \frac{1}{2}\mathrm{Spark}(\Phi)$, then x is the unique solution to

$$\operatorname*{argmin}_x \sum_{i=1}^{m} I(x_i \neq 0) \quad \text{subject to} \quad \Phi x = y$$

Mutual Coherence

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

$$\mu(\Phi) \triangleq \max_{1 \le i, j \le m; i \ne j} \frac{\left| \phi_i^T \phi_j \right|}{\left\| \phi_i \right\|_2 \left\| \phi_j \right\|_2}$$

Performances of ℓ_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

$$\underset{x}{\operatorname{argmin}} \sum_{i=1}^{m} |x_i| \quad \text{subject to} \quad \Phi x = y$$

Performances of ℓ_1 diversity minimization algorithms

- Noisy Case [Donoho et al 06]

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

Then the solution

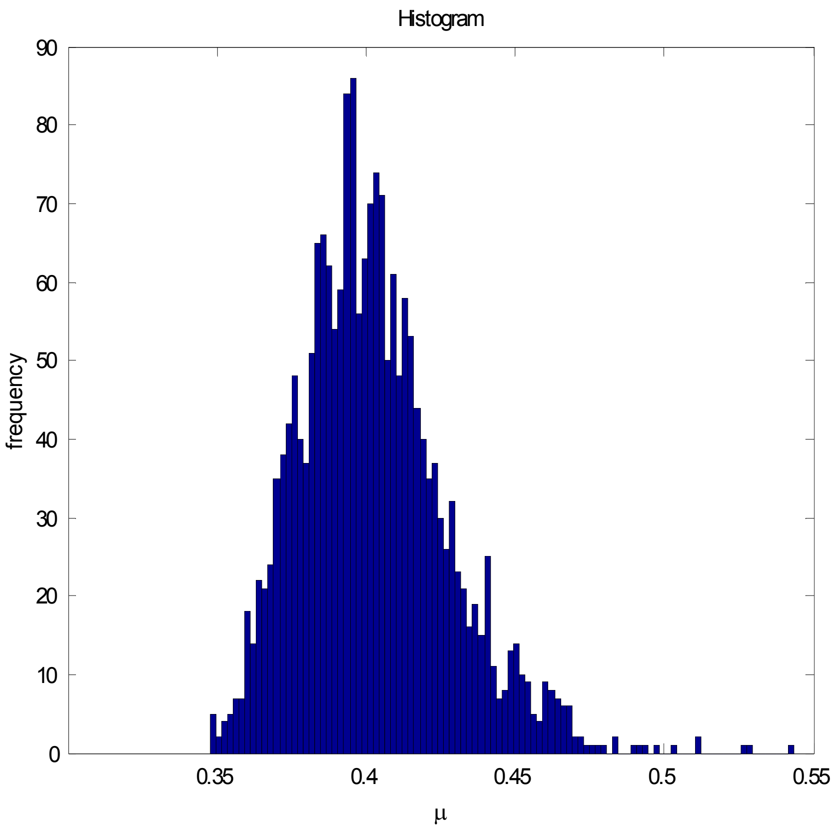
$$d^* = \arg\min_d \sum_{i=1}^{m} |d_i| \quad \text{subject to} \quad \|y - \Phi d\|_2 \le \beta$$

satisfies

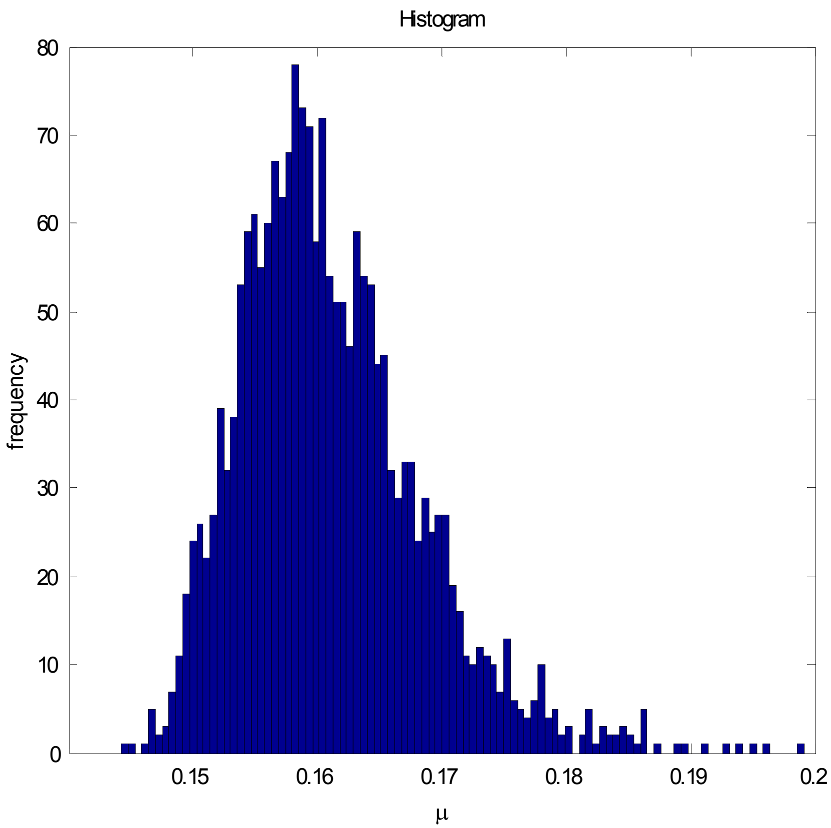
$$\|d^* - x\|_2^2 \le \frac{4\beta^2}{1 - \mu(\Phi)\left(4\|x\|_0 - 1\right)}$$

Empirical distribution of mutual coherence (Gaussian matrices)

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



$\Phi_{100\times200}$, **mean = 0.4025, std = 0.0249**



$\Phi_{1000\times2000}$, **mean = 0.161, std = 0.0073**

Performance of Orthogonal Matching Pursuit

- Noiseless Case [Tropp 04]

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

Performance of Orthogonal Matching Pursuit

- Noisy Case [Donoho et al]

  Assume $y = \Phi x + \varepsilon$, $\|\varepsilon\|_2 \le \beta$, $x_{\min} = \min_{1 \le i \le m} |x_i|$,

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

  Stop OMP when residual error $\le \beta$.

  Then the solution of OMP satisfies

  $$\|x_{OMP} - x\|_2^2 \le \frac{\beta^2}{1 - \mu(\Phi)\left(\|x\|_0 - 1\right)}$$

Restricted Isometry Constant

- Definition [Candes et al]

For a matrix Φ, the smallest constant δ_k such that

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

for any k-sparse signal x.

Performances of ℓ_1 diversity measure minimization algorithms

- [Candes 08]

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

Then, the solution

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

satisfies

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

where C only depends on δ_{2k}.

Performances of ℓ_1 diversity measure minimization algorithms

- [Candes 08]

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

Then, the solution

$$d^* = \underset{d}{\operatorname{argmin}} \sum_{i=1}^{m} |d_i| \quad \text{subject to} \quad \|y - \Phi d\|_2 \le \beta$$

satisfies

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

where C_1, C_2 only depend on δ_{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 ℓ_1 based procedure, number of measurements required are $n \gtrsim k \log m$

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: $y = \Phi x + \varepsilon$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, i.i.d.
- Random matrix Φ, $\phi_{ij} \sim N(0, \sigma_a^2)$, i.i.d.
- Performance metric: exact support recovery
- Consider a sequence of problems with increasing sizes
- $k = 2$: $c(x) \triangleq \min_{T \subseteq \{1,2\}} \left\{ \frac{1}{2|T|} \log\left(1 + \frac{\sigma_a^2}{\sigma_\varepsilon^2} \sum_{i \in T} x_{nonzero,i}^2 \right) \right\}$. (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

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

Necessary condition

If there exists a sequence of support recovery methods such that

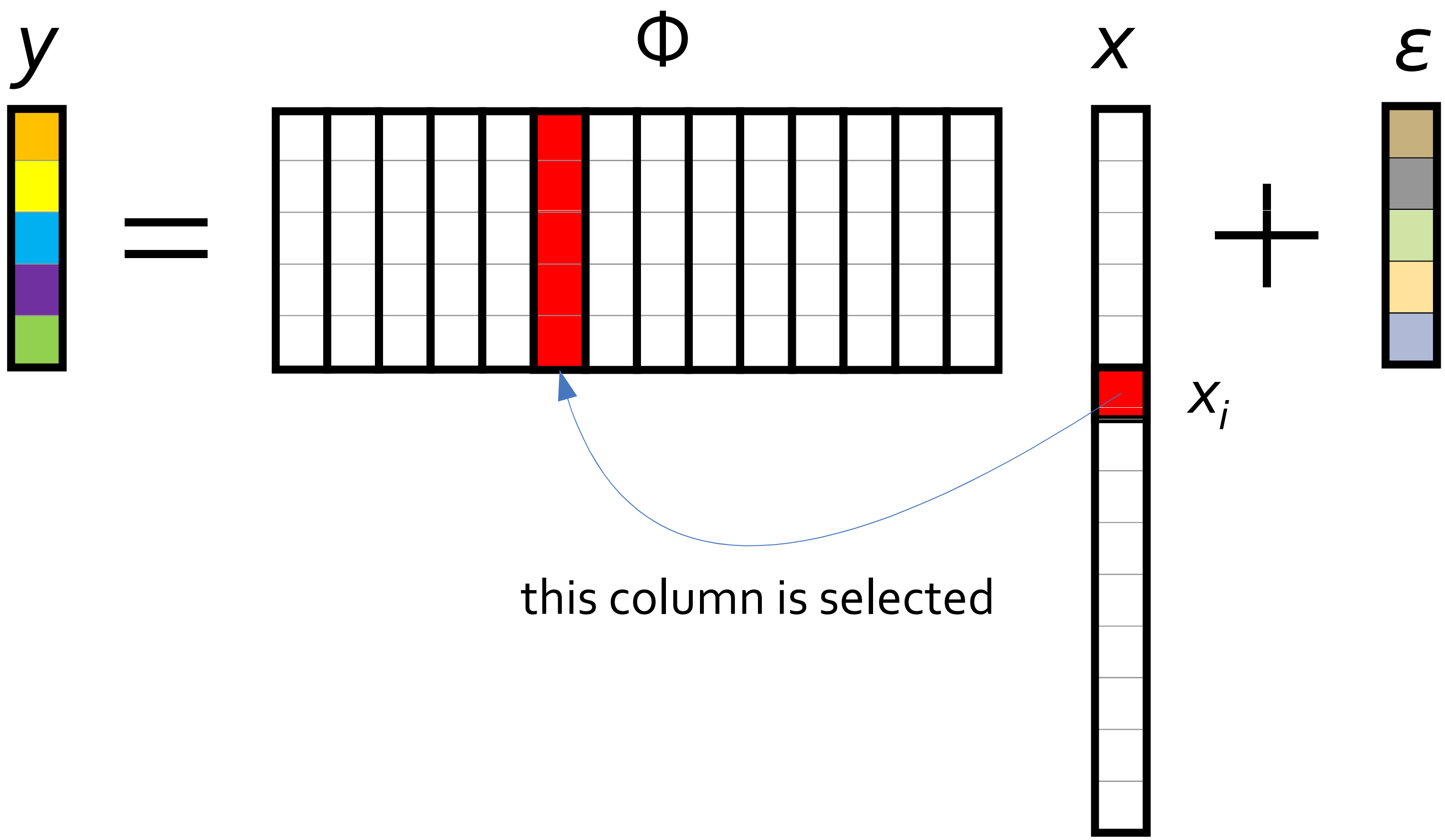
$$\lim_{m \to \infty} P\{\mathrm{supp}(\hat{x}) \neq \mathrm{supp}(x)\} = 0$$

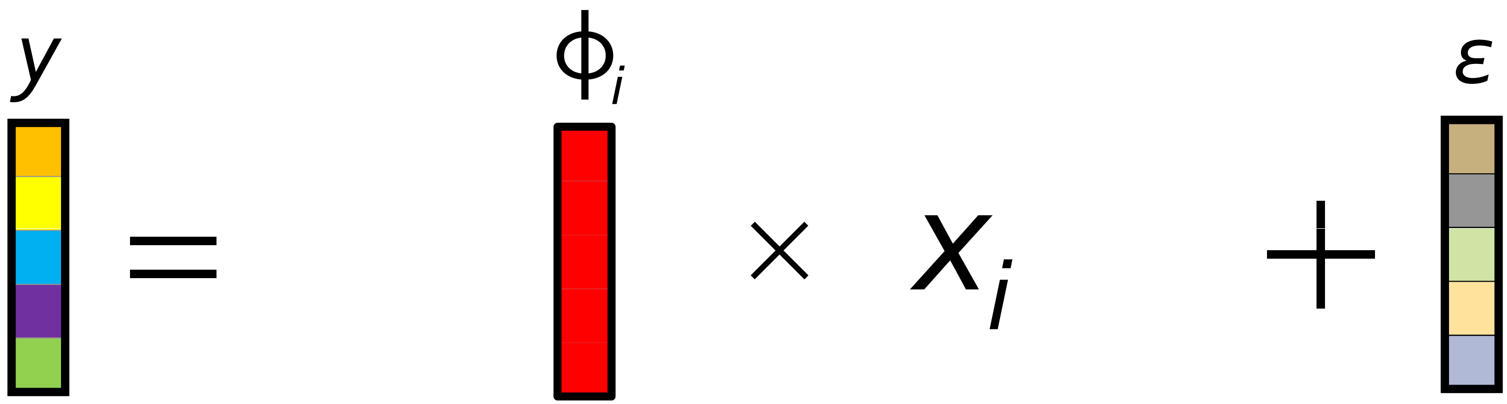
then

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

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

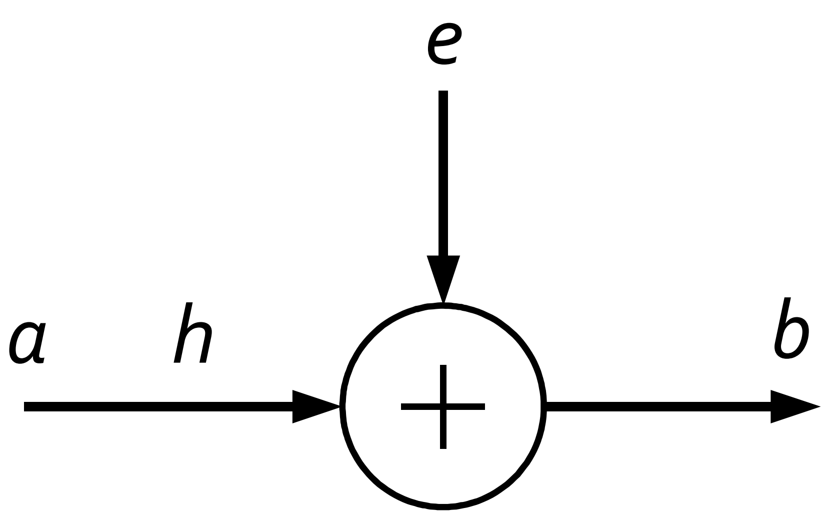
Network information theory perspective
Connection to channel coding (K = 1)





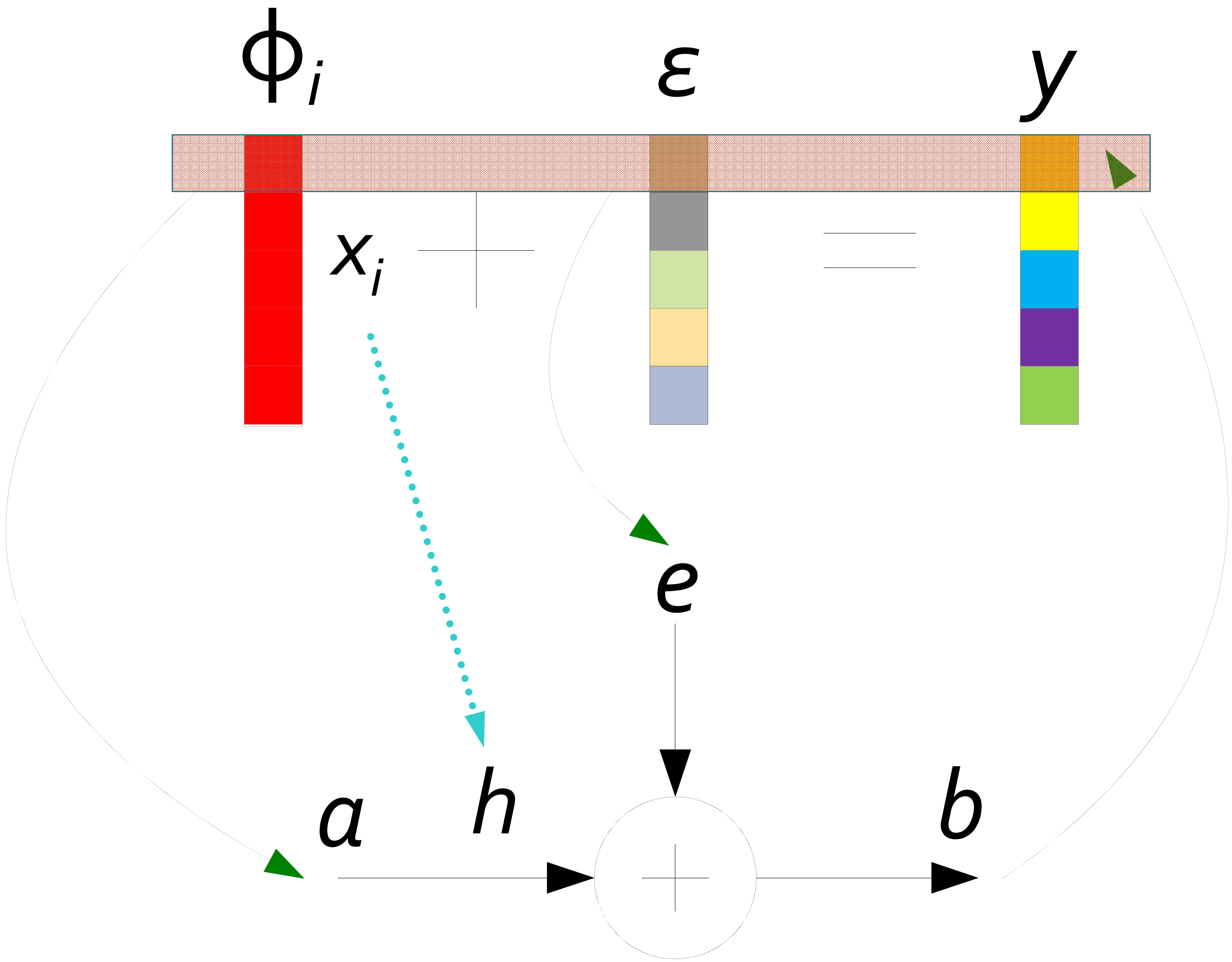
$$y = x_i \phi_i + \varepsilon$$

Additive White Gaussian Noise (AWGN) Channel

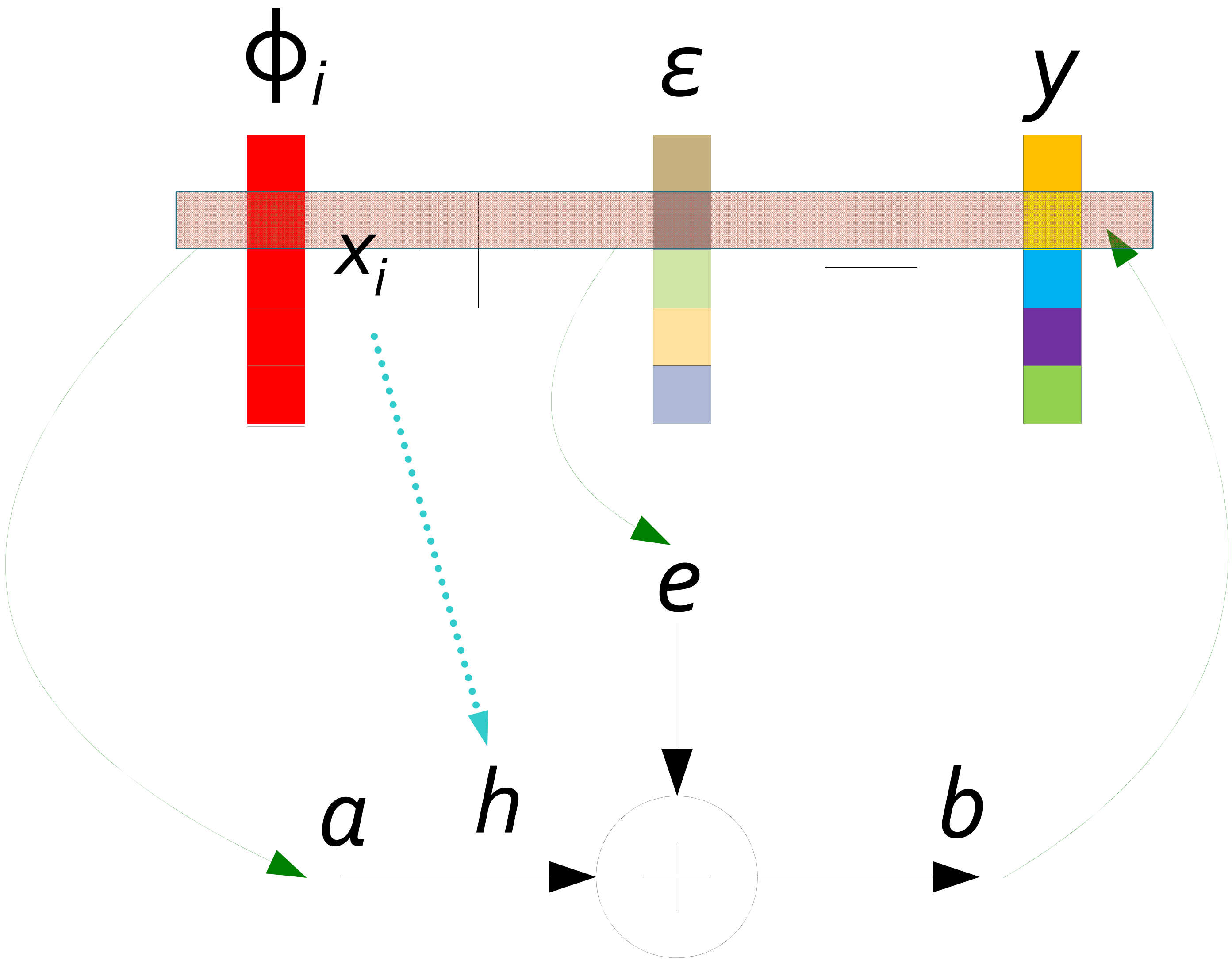


- a: channel input
- h: channel gain
- e: additive Gaussian noise
- b: channel output, $b = ha + e$

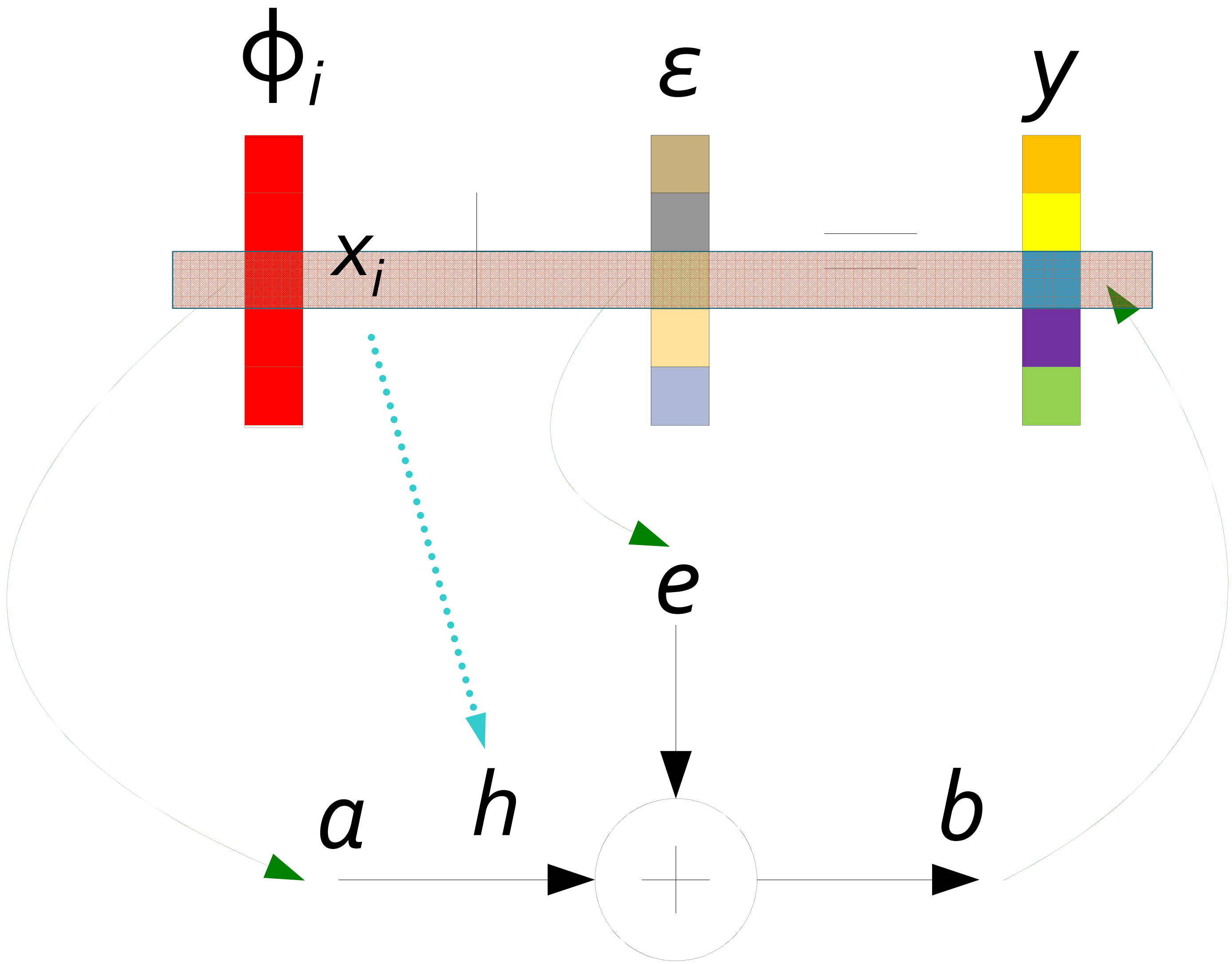
Recall
$y = x_i \phi_i + \varepsilon$

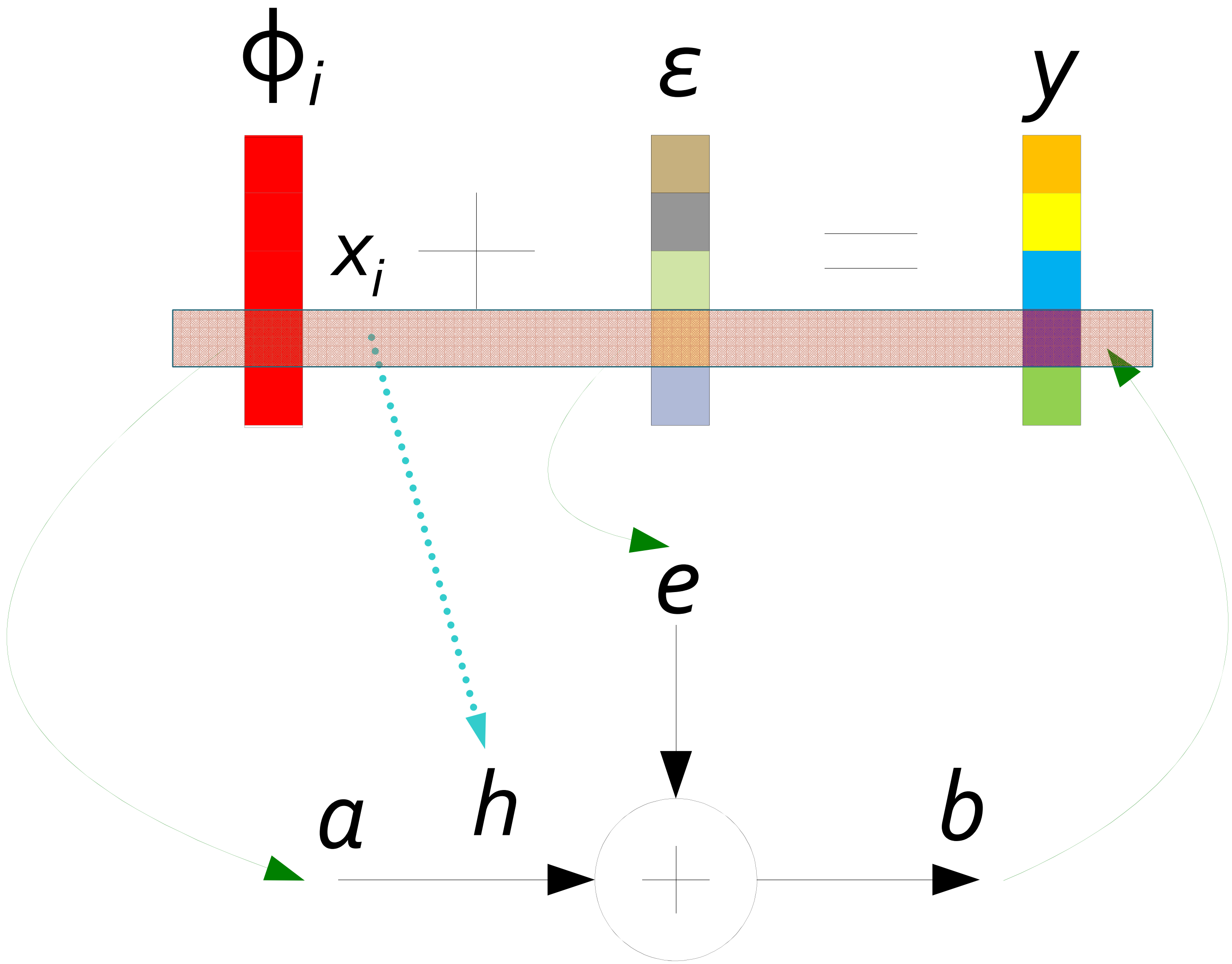


φi
ε
y
xi
e
a
h
b

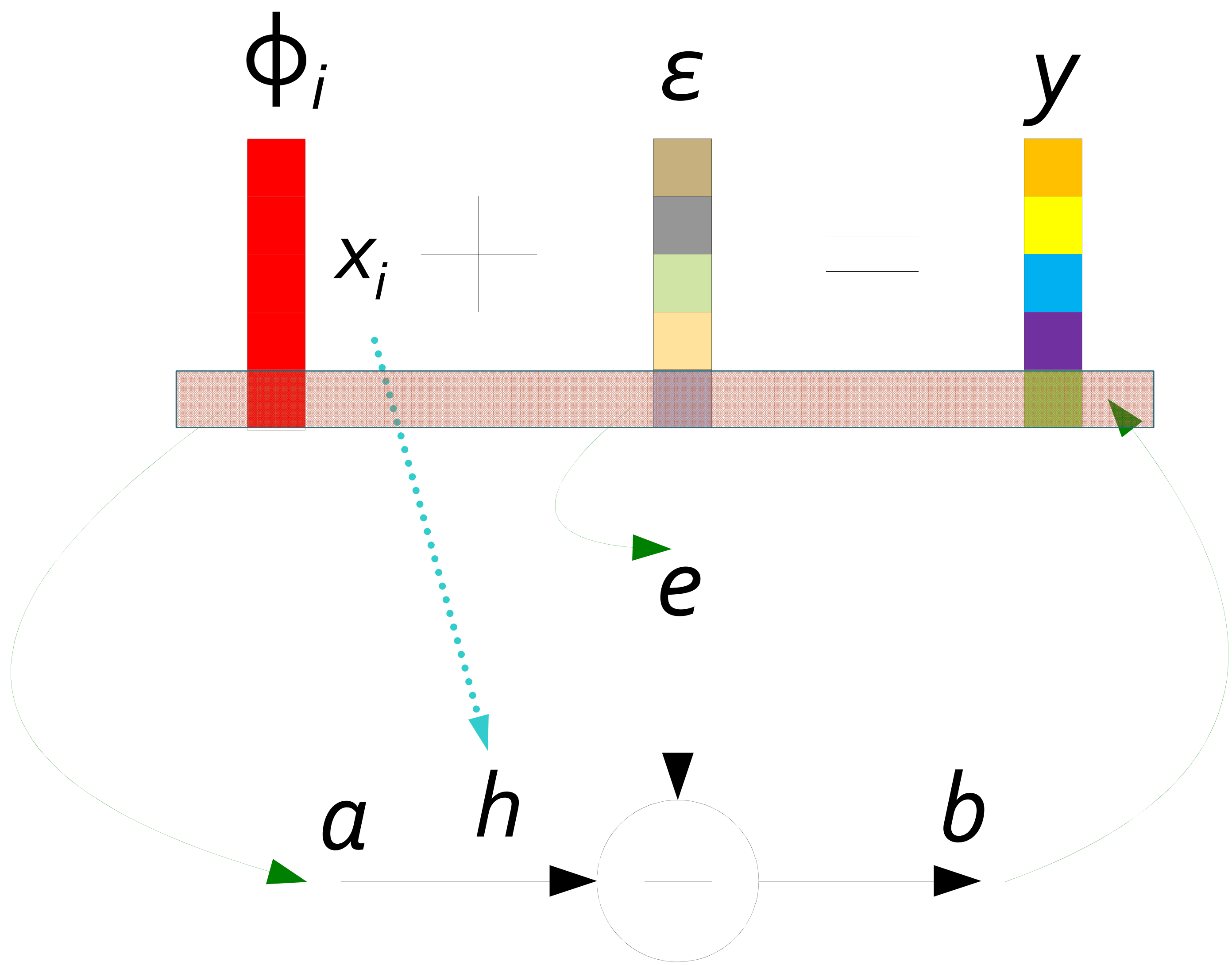


ϕi
ε
y
xi
e
a
h
b


φi
ε
y
xi
e
a
h
b

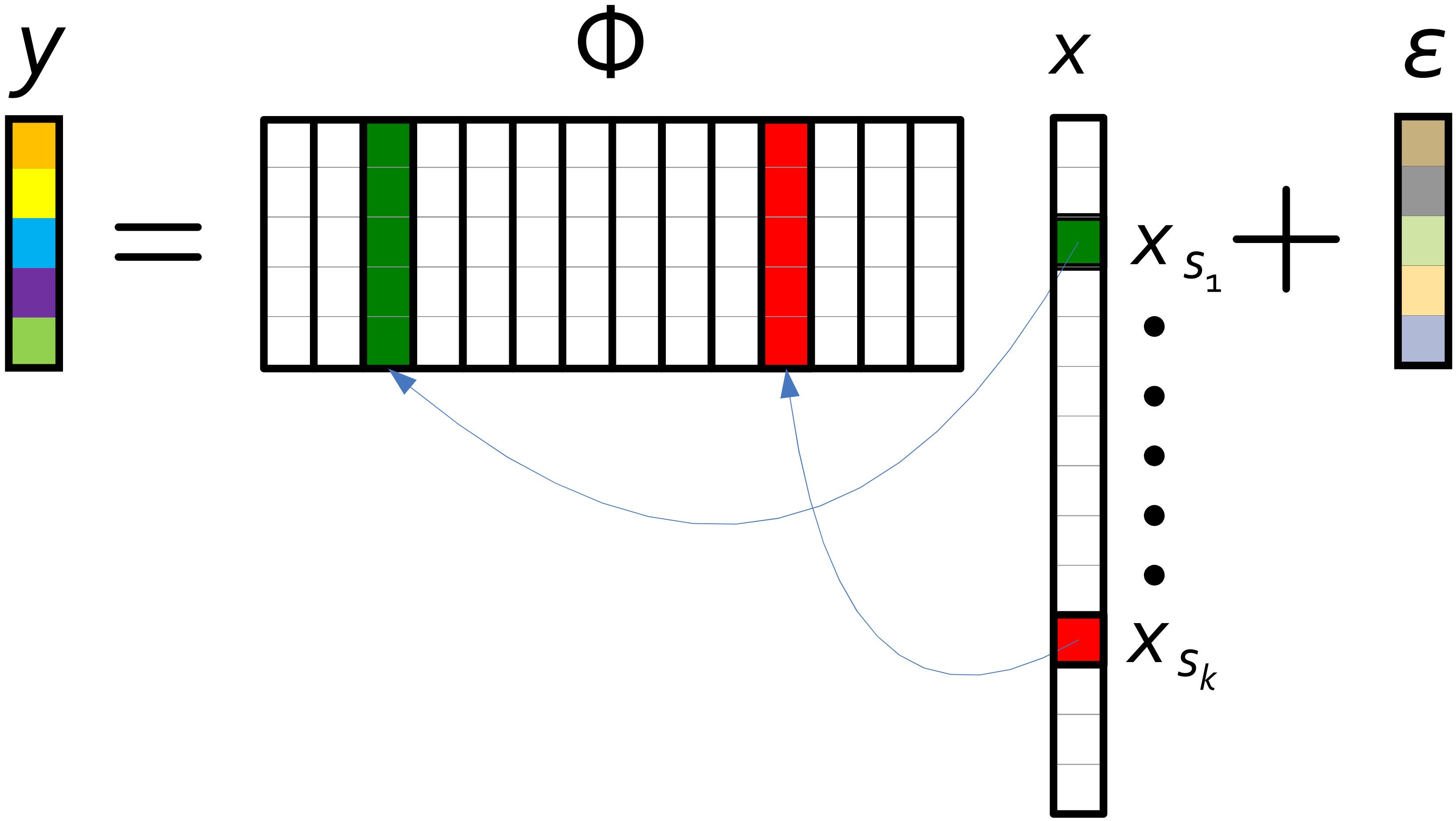


ϕ_i
x_i
ε
y
e
a
h
b

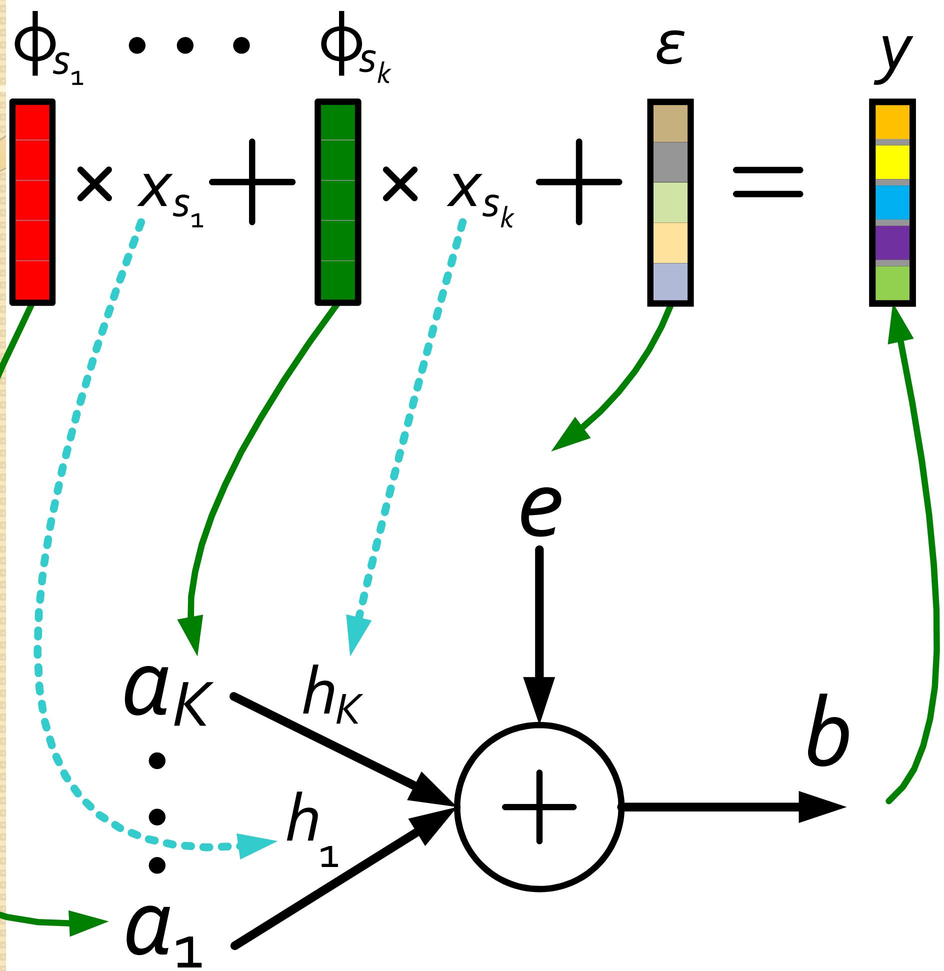


ϕ_i
x_i
+
ε
=
y
e
a
h
b

Connection to channel coding: ($K \geq 1$)



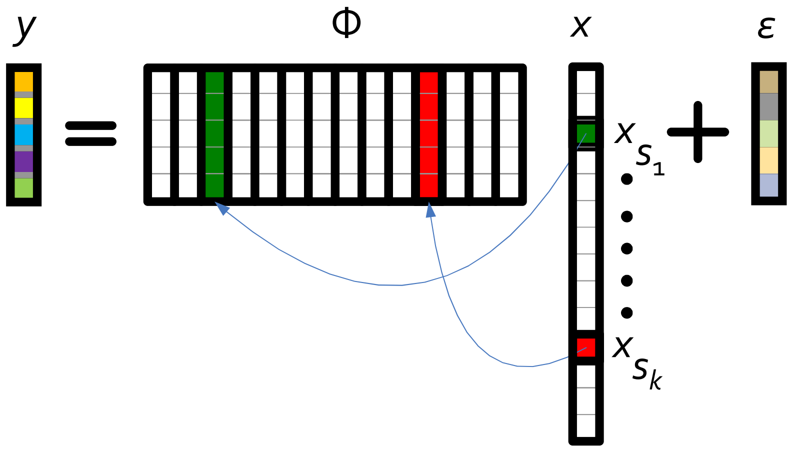
Multiple Access Channel (MAC)

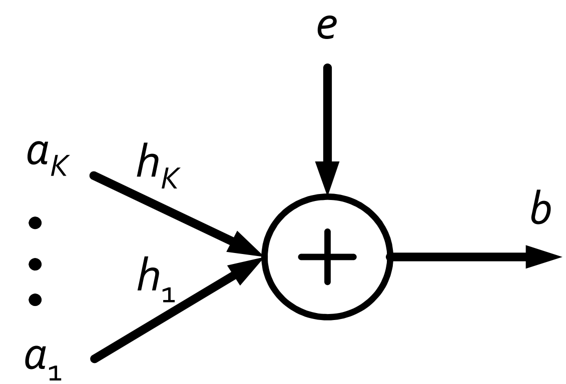


$$y = x_{s1}\phi_{s1} + \ldots + x_{sk}\phi_{sk} + \varepsilon$$

$$b = h_1 a_1 + \ldots + h_K a_K + e$$

Connection between two problems





Φ: dictionary, measurement matrix	codebook
ϕ_i: a column	a codeword
m different positions in x	message set {1, 2, …, *m*}
$s_1,\ldots,s_k$: indices of nonzeros	the messages selected
x_{s_i}: source activity	channel gain h_i

Differences between two problems

Channel coding	Support recovery
Each sender works with a codebook designed <u>only</u> for that sender.	All "senders" share the <u>same</u> codebook A. Different senders select different codewords. (**Common codebook**)
Capacity result available when channel gain h_i is <u>known</u> at receiver.	We <u>don't</u> know the nonzero signal activities x_i. (**Unknown channel**)

- 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 ℓ_1 minimization (e.g., BP), place heavy restrictions on the form of the dictionary Φ.

- 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 $\mathbf{x}_0$ can greatly affect the difficulty of canonical sparse recovery problems.

- This effect is strongly algorithm-dependent …

Examples:
ℓ_1 -norm Solutions and OMP

- With *ℓ_1 -norm solutions,* performance is independent of the magnitudes of nonzero coefficients [Malioutov et al., 2004].

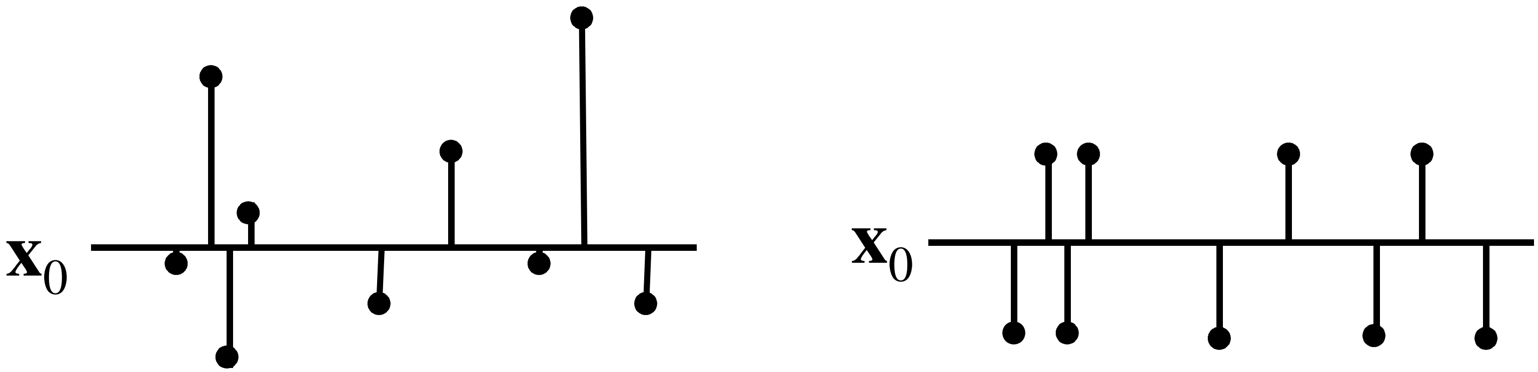
- **Problem**: Performance does not improve when the situation is favorable.

- *OMP* is highly sensitive to these magnitudes.

- **Problem**: Perform degrades heavily with unit magnitudes.

In General

- If the magnitudes of the non-zero elements in $\mathbf{x}_0$ are highly scaled, then the canonical sparse recovery problem should be easier.

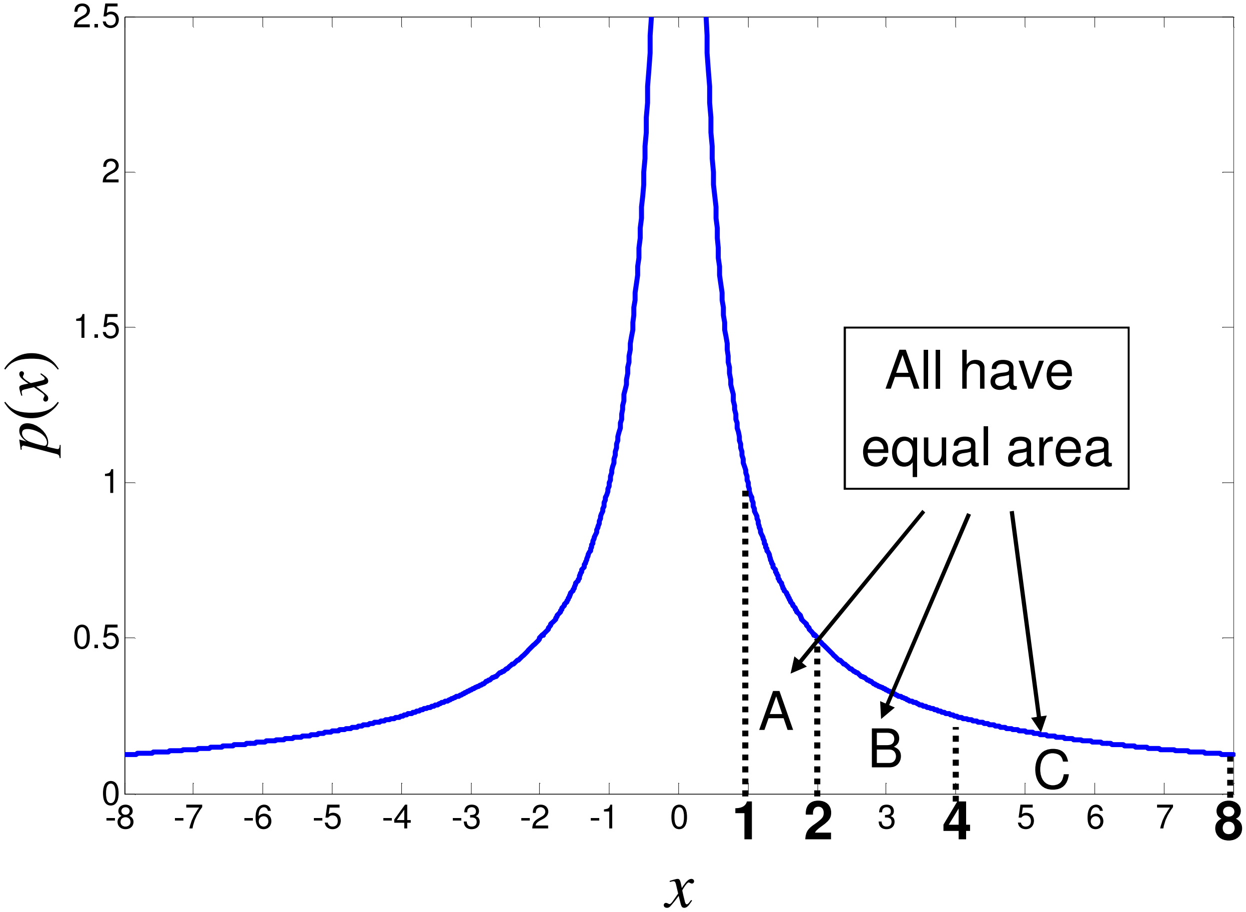


scaled coefficients (easy) uniform coefficients (hard)

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

Jeffreys Distribution

$$\text{Density:} \quad p(x) \propto \frac{1}{|x|}$$



Empirical Example

- For each test case:

  1. Generate a random dictionary Φ with ***50 rows*** and ***100 columns***.

  2. Generate a sparse coefficient vector $\mathbf{x}_0$.

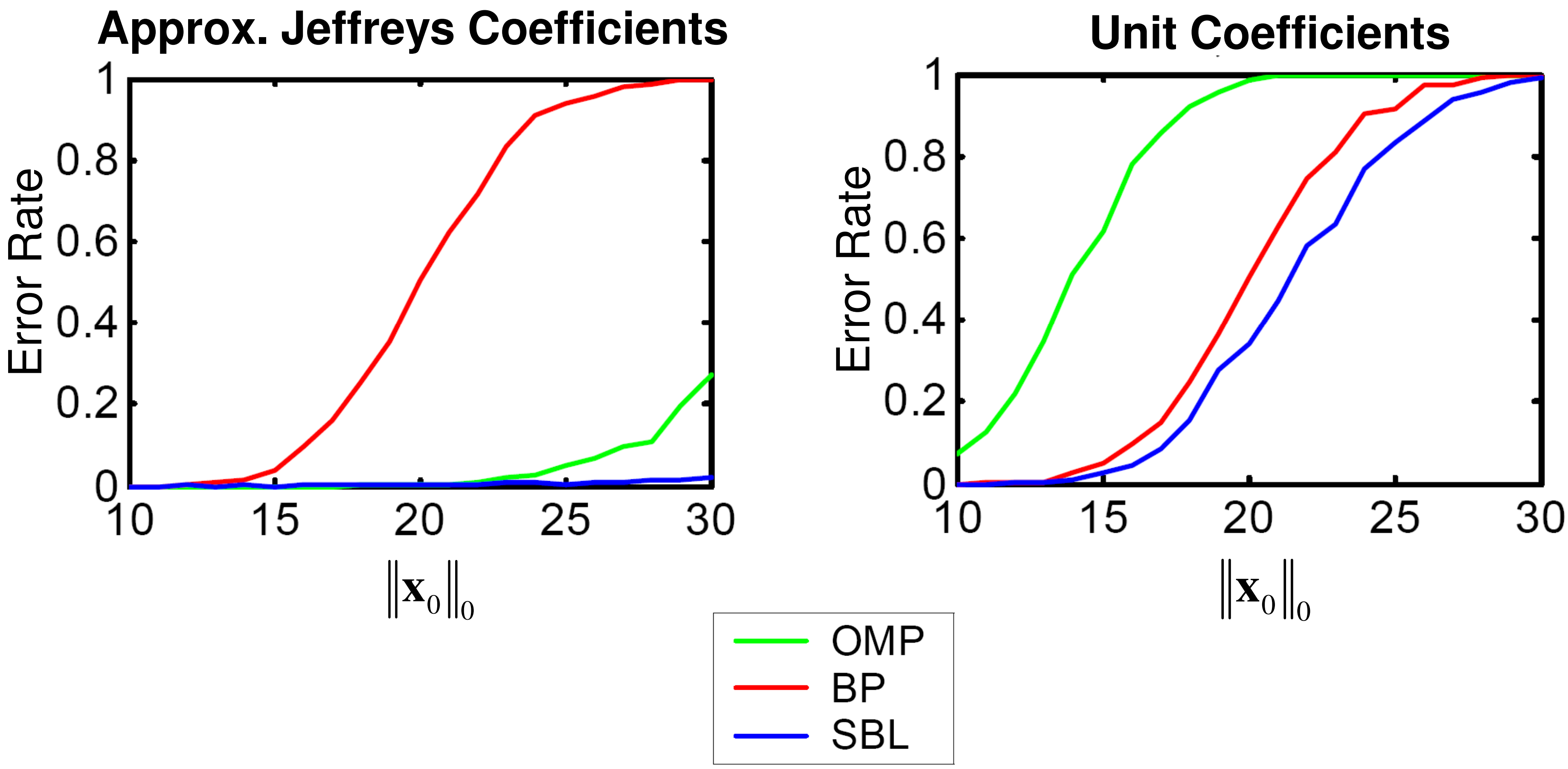
  3. Compute signal via $\mathbf{y} = \Phi\, \mathbf{x}_0$ (noiseless).

  4. Run ***BP*** and ***OMP***, as well as a competing Bayesian method called ***SBL*** (more on this later) to try and correctly estimate $\mathbf{x}_0$.

  5. Average over1000 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 Φ.
- 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.

<u>Section II</u>:

MAP Estimation Using the Sparse Linear Model

Overview

- Can be viewed as sparse penalized regression with general sparsity-promoting penalties (ℓ_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 ℓ_1 solution on sparse recovery problems.

Sparse Linear Model

- Linear generative model:

$$\mathbf{y} = \Phi\mathbf{x} + \boldsymbol{\varepsilon}$$

Observed n-dimensional data vector

Matrix of m basis vectors

Unknown Coefficients

Gaussian noise with variance λ

- **Objective**: Estimate the unknown $\mathbf{x}$ given the following assumptions:

1. Φ is *overcomplete*, meaning the number of columns m is greater than the signal dimension n.

2. $\mathbf{x}$ is *maximally sparse*, i.e., many elements equal zero.

Sparse Inverse Problem

- Noiseless case ($\varepsilon = 0$):

$$\mathbf{x}_0 \triangleq \arg\min_{\mathbf{x}} \underbrace{\|\mathbf{x}\|_0}_{\boxed{\ell_0 \text{ norm } = \text{ \# of nonzeros in } \mathbf{x}}} \quad \text{s.t.} \quad \mathbf{y} = \Phi\mathbf{x}$$

- Noisy case ($\varepsilon > 0$):

$$\mathbf{x}_0(\lambda) \triangleq \arg\min_{\mathbf{x}} \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0$$

$$= \arg\max_{\mathbf{x}} \underbrace{\exp\left[-\tfrac{1}{2\lambda}\|\mathbf{y} - \Phi\mathbf{x}\|_2^2\right]}_{\text{likelihood}} \underbrace{\exp\left[-\tfrac{1}{2}\|\mathbf{x}\|_0\right]}_{\text{prior}}$$

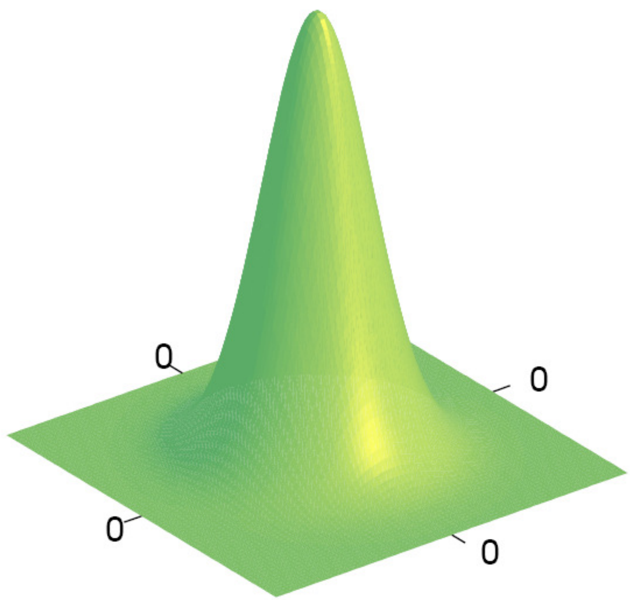
Difficulties

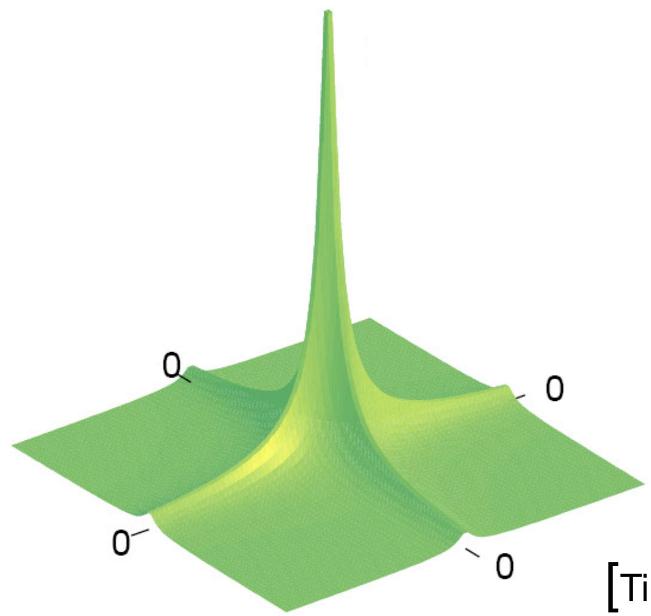
1. Combinatorial number of local minima

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





[Tipping, 2001]

Gaussian Distribution

Sparse Distribution

Basic MAP Estimation

$$\begin{aligned}\hat{\mathbf{x}} &\triangleq \arg\max_{\mathbf{x}} \; p(\mathbf{x} \mid \mathbf{y}) \\ &= \arg\min_{\mathbf{x}} \; -\log p(\mathbf{y} \mid \mathbf{x}) - \log p(\mathbf{x}) \\ &= \arg\min_{\mathbf{x}} \; \underbrace{\|\mathbf{y} - \Phi\mathbf{x}\|_2^2}_{\text{data fit}} + \lambda \underbrace{\sum_{i=1}^{n} g(x_i)}_{}\end{aligned}$$

data fit

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

Note: Bayesian interpretation will be useful later …

Example Sparsity Penalties

- With $g(x_i) = \mathrm{I}[x_i \neq 0]$ we have the canonical sparsity penalty and its associated problems.

- Practical selections:

$$g(x_i) = \log(x_i^2 + \varepsilon), \quad \text{[Chartrand and Yin, 2008]}$$

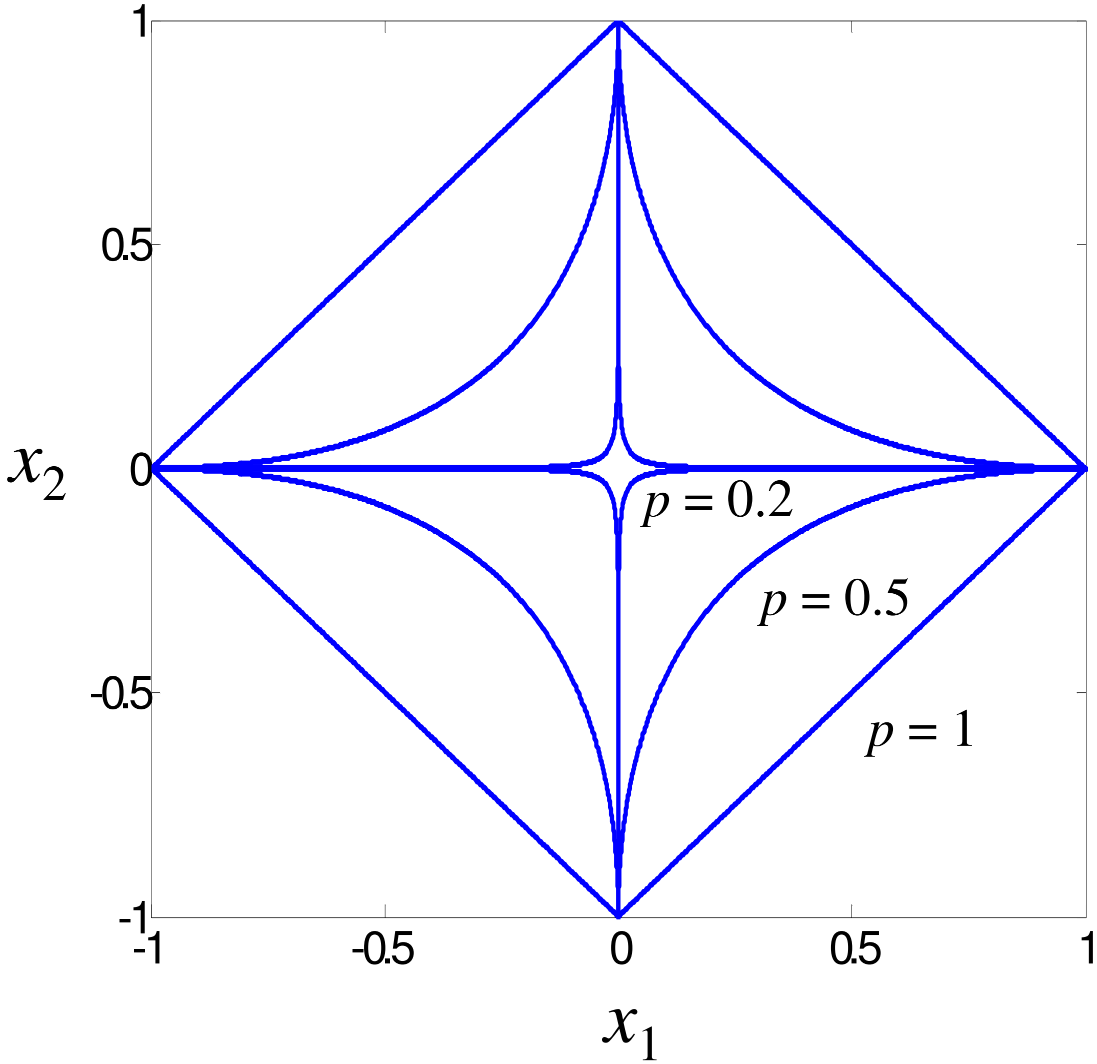
$$g(x_i) = \log(|x_i| + \varepsilon), \quad \text{[Candes et al., 2008]}$$

$$g(x_i) = |x_i|^p, \quad \text{[Rao et al., 2003]}$$

- Different choices favor different levels of sparsity.

Example 2-D Contours

$$g(x_i) = |x_i|^p$$



Which Penalty Should We Choose?

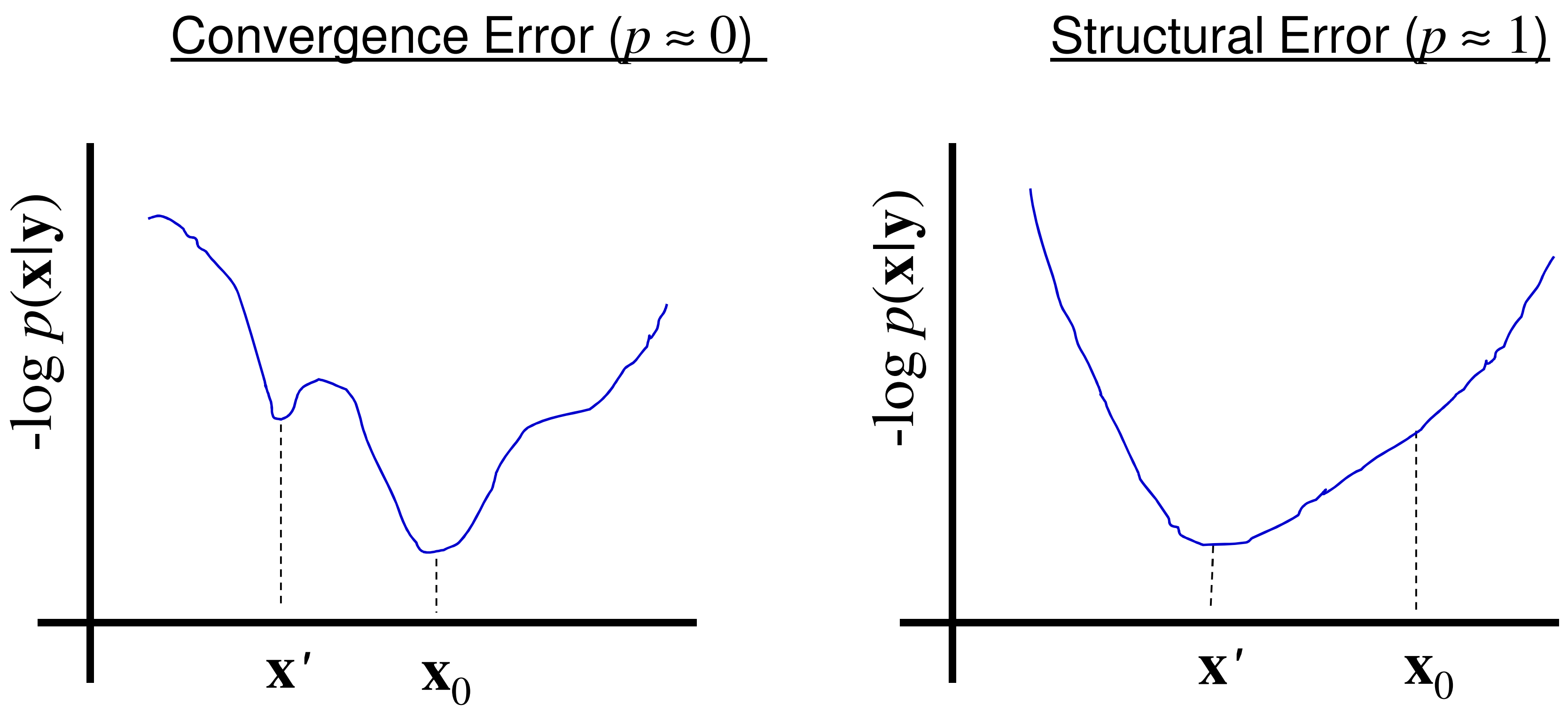
- Two competing issues:

  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 $\mathbf{x}_0$: structural error.

- Answer is ultimately application- and algorithm-dependent

Convergence Errors vs. Structural Errors



$\mathbf{x}'$ = solution we have converged to

$\mathbf{x}_0$ = maximally sparse solution

Desirable Properties of Algorithms

- Can be implemented via relatively simple primitives already in use, e.g., ℓ_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

$$\mathrm{Y} = \Phi\,\mathrm{X} + \mathrm{E}$$

- **Optimization Problem**:

$$\mathrm{X}_0(\lambda) \;=\; \arg\min_{\mathrm{X}} \;\left\|\mathrm{Y} - \Phi\,\mathrm{X}\right\|_F^2 + \lambda \underbrace{\sum_{i=1}^{m} \mathrm{I}\left[\left\|\mathbf{x}_{i\cdot}\right\| \neq 0\right]}_{\boxed{\text{\# of nonzero rows in } \mathrm{X}}}$$

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

Reweighted ℓ_2 Optimization

- Assume: $g(x_i) = h(x_i^2), \quad h$ concave
- Updates:

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

$$= \tilde{W}^{(k)} \Phi^T \left(\lambda I + \Phi \tilde{W}^{(k)} \Phi^T\right)^{-1} \mathbf{y}$$

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

- 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: $g(x_i) = |x_i|^p, \;\; 0 \le p \le 2$
- 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: $g(x_i) = \log\left(x_i^2 + \varepsilon\right), \;\; \varepsilon \ge 0$
- Weight update: $w_i^{(k+1)} \rightarrow \left[\left(x_i^{(k+1)}\right)^2 + \varepsilon\right]^{-1}$
- Properties: Slowly reducing ε to zero smoothes out local minima initially allowing better solutions to be found; very useful for recovering scaled coefficients …

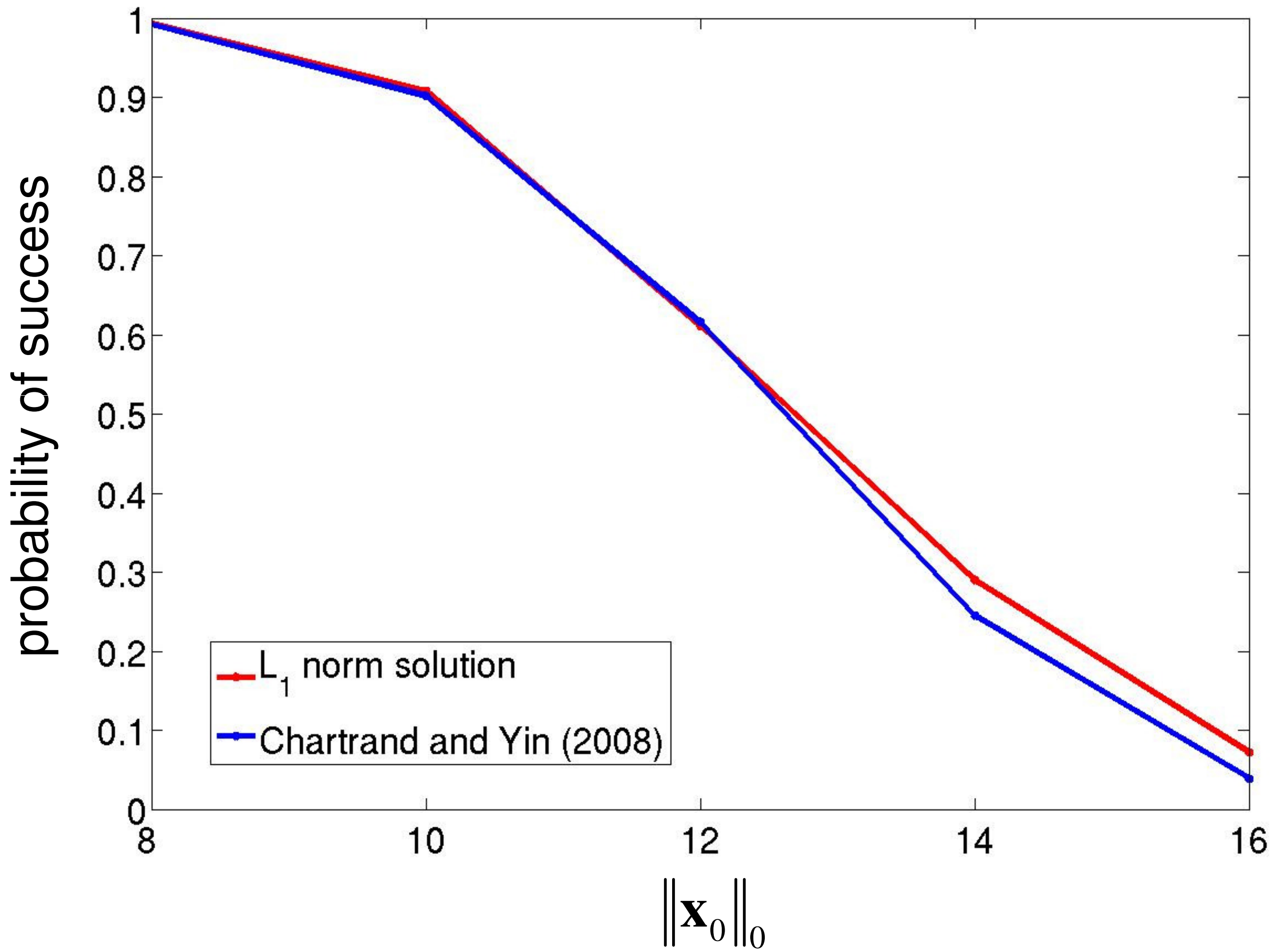
Empirical Comparison

- For each test case:

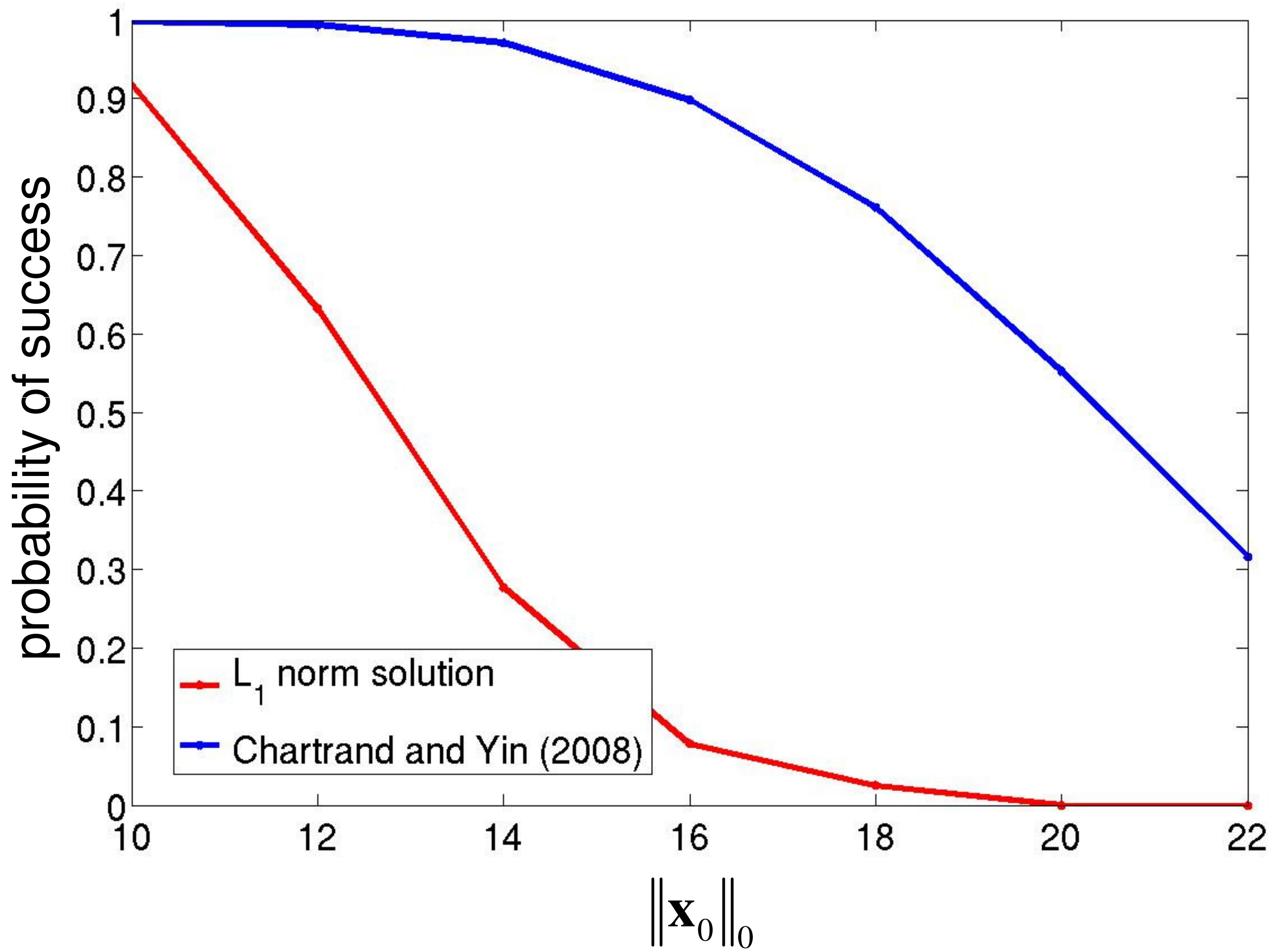
  1. Generate a random dictionary Φ with *50 rows* and *250 columns*
  2. Generate a sparse coefficient vector $\mathbf{x}_0$.
  3. Compute signal via $\mathbf{y} = \Phi\, \mathbf{x}_0$ (noiseless).
  4. Compare *Chartrand and Yin's reweighted ℓ_2 method* with ℓ_1 norm solution with regard to estimating $\mathbf{x}_0$.
  5. Average over 1000 independent trials.

- Repeat with different sparsity levels and different nonzero coefficient distributions.

Empirical: Unit Nonzeros



Results: Gaussian Nonzeros



Reweighted ℓ_1 Optimization

- Assume: $$g(x_i) = h(|x_i|), \quad h \text{ concave}$$

- Updates: $$\mathbf{x}^{(k+1)} \rightarrow \arg\min_{\mathbf{x}} \|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda \sum_i w_i^{(k)} |x_i|$$ $$w_i^{(k+1)} \rightarrow \left. \frac{\partial g(x_i)}{\partial |x_i|} \right|_{x_i = x_i^{(k+1)}}$$

- 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 $\mathbf{x}$.

Example [Candes et al., 2008]

- Penalty: $$g(x_i) = \log\left(|x_i| + \varepsilon\right), \quad \varepsilon \geq 0$$

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

- When nonzeros in $\mathbf{x}_0$ are scaled, works much better than regular ℓ_1, depending on how ε is chosen.

- Local minima exist, but since each iteration is sparse, local solutions are not so bad (no worse than regular ℓ_1 solution).

Empirical Comparison

- For each test case:

1. Generate a random dictionary Φ with *50 rows* and *100 columns.*

2. Generate a sparse coefficient vector $\mathbf{x}_0$ with *30* truncated Gaussian, strictly positive nonzero coefficients.

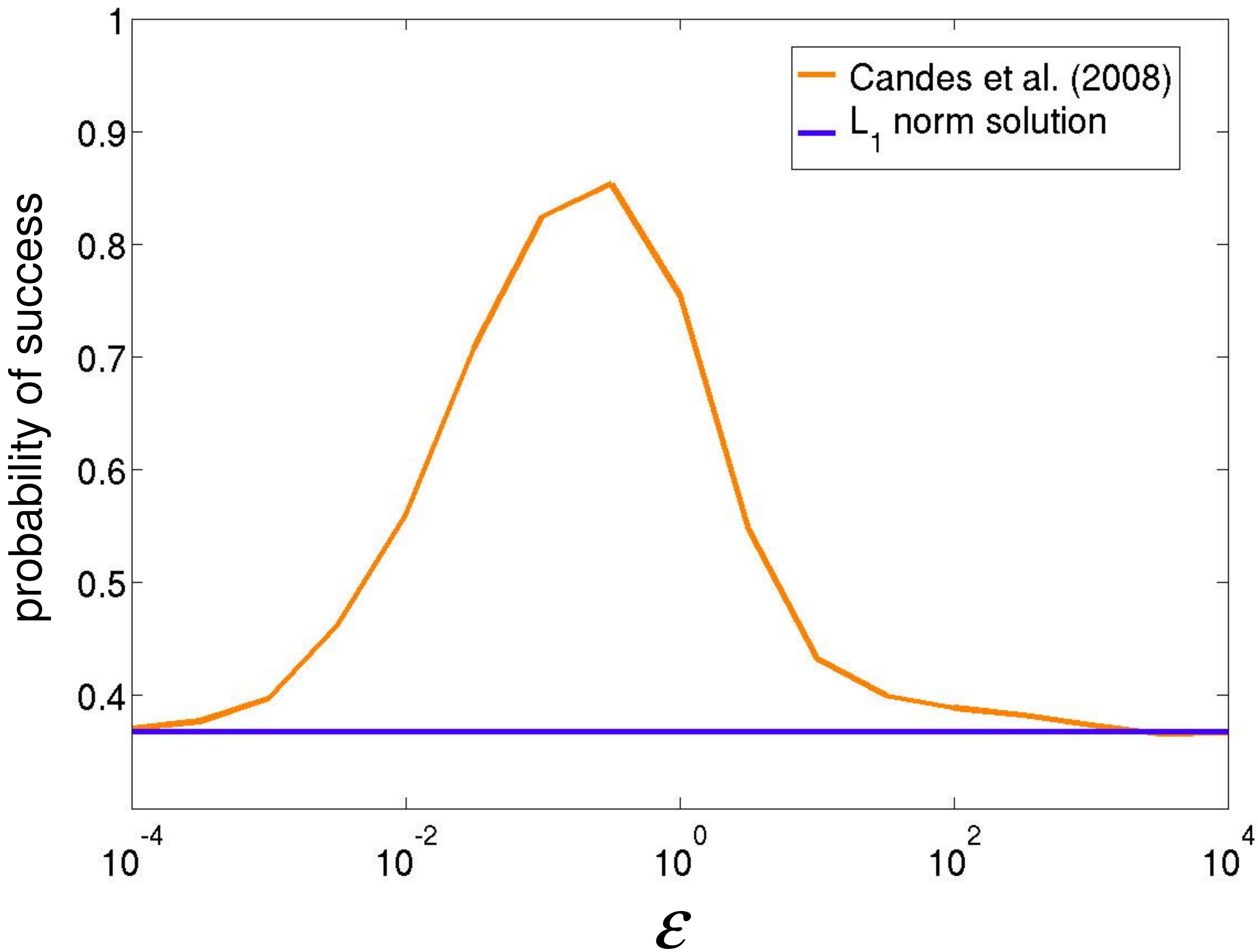
3. Compute signal via $\mathbf{y} = \Phi\,\mathbf{x}_0$ (noiseless).

4. Compare *Candes et al.'s reweighted ℓ_1 method* (10 iterations) with *ℓ_1 norm solution*, both constrained to be *non-negative* to try and estimate $\mathbf{x}_0$.

5. Average over 1000 independent trials.

- Repeat with different values of the parameter ε .

Empirical Comparison



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 $p(\mathbf{x}|\mathbf{y})$ by maximizing the joint distribution

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})$$

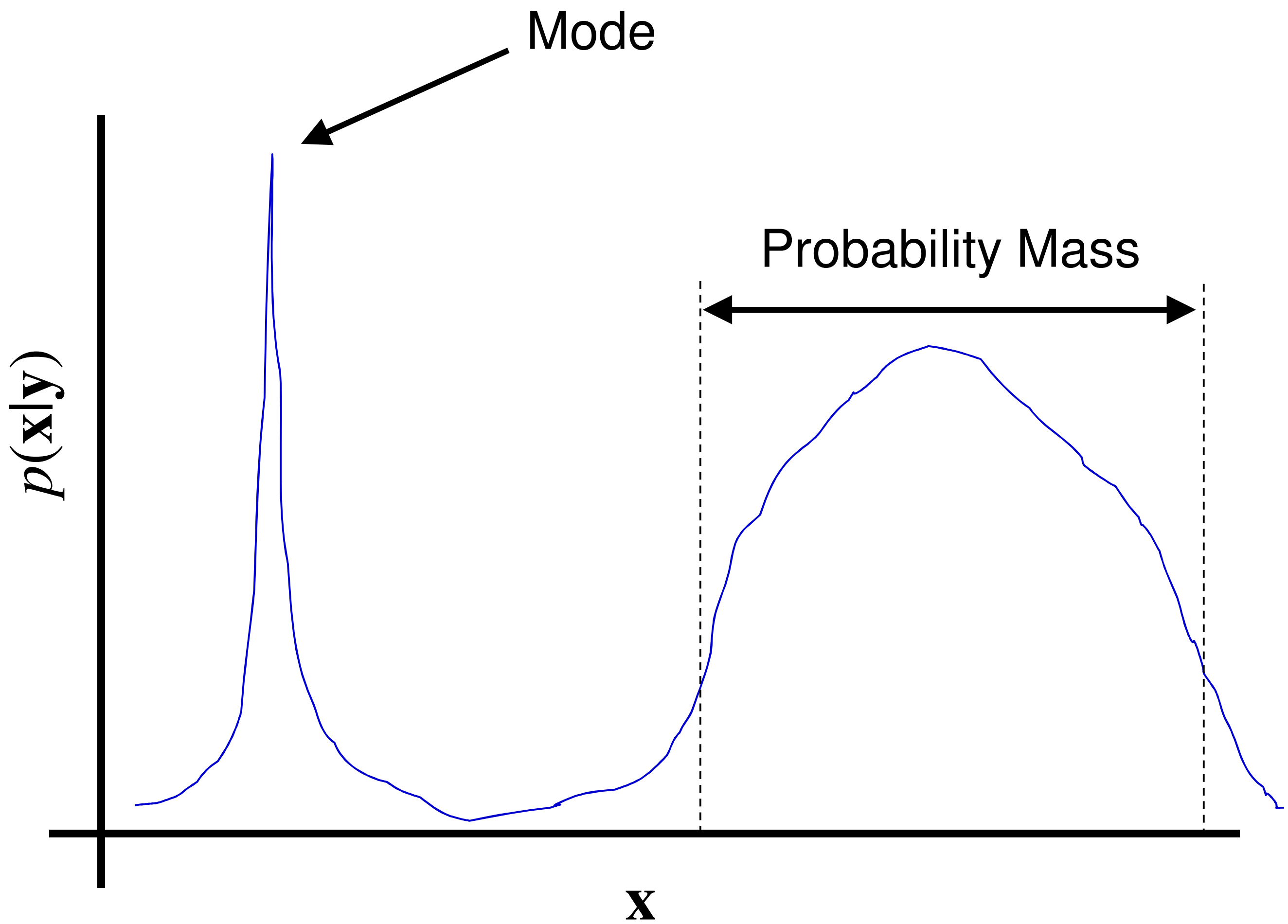
- ***Bayesian inference*** uses posterior information beyond the mode, i.e., posterior *mass*:

- 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



Problem

- For essentially all sparse priors, cannot compute normalized posterior

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})}{\int p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}}$$

- Also cannot computer posterior moments, e.g.,

$$\begin{aligned} \propto_x &= \mathrm{E}[\mathbf{x} \mid \mathbf{y}] \\ \Sigma_x &= \mathrm{Cov}[\mathbf{x} \mid \mathbf{y}] \end{aligned}$$

- So efficient approximations are needed …

Approximating Posterior Mass

- **Goal**: Approximate $p(\mathbf{x},\mathbf{y})$ with some distribution $\hat{p}(\mathbf{x},\mathbf{y})$ that

  1. Reflects the significant mass in $p(\mathbf{x},\mathbf{y})$.
  2. Can be normalized to get the posterior $p(\mathbf{x}|\mathbf{y})$.
  3. Has easily computable moments, e.g., can compute $\mathrm{E}[\mathbf{x}|\mathbf{y}]$ or $\mathrm{Cov}[\mathbf{x}|\mathbf{y}]$.

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

$$\int \left| p(\mathbf{x},\mathbf{y}) - \hat{p}(\mathbf{x},\mathbf{y}) \right| d\mathbf{x} \quad = \quad \int p(\mathbf{y} \mid \mathbf{x}) \left| p(\mathbf{x}) - \hat{p}(\mathbf{x}) \right| d\mathbf{x}$$

Recipe

1. Start with a Gaussian likelihood

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

2. Pick an appropriate prior $p(\mathbf{x})$ that encourages sparsity

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

4. Solve: $\hat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \int p(\mathbf{y} \mid \mathbf{x}) \left| p(\mathbf{x}) - p(\mathbf{x};\boldsymbol{\gamma}) \right| d\mathbf{x}$

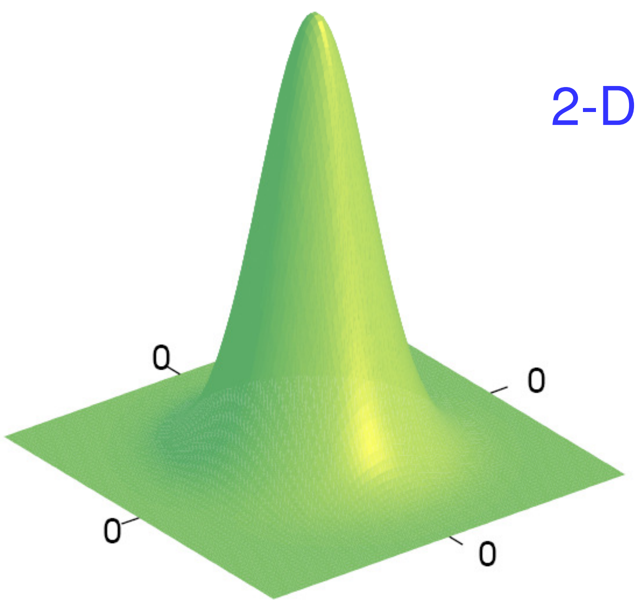
5. Normalize: $p(\mathbf{x} \mid \mathbf{y};\hat{\boldsymbol{\gamma}}) = \dfrac{p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x};\hat{\boldsymbol{\gamma}})}{\int p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x};\hat{\boldsymbol{\gamma}})\, d\mathbf{x}}$

Step 2: Prior Selection

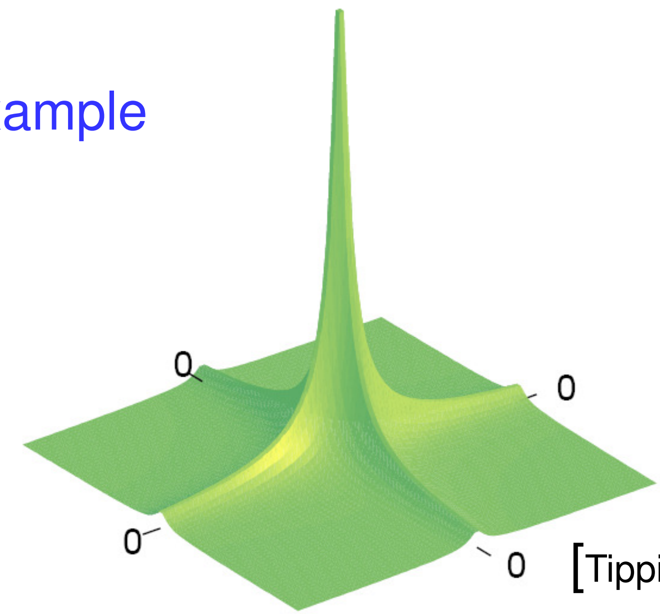
- Assume a sparse prior distribution on each coefficient:

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

2-D example



Gaussian Distribution



[Tipping, 2001]

Sparse Distribution

Step 3: Forming Approximate Priors $p(\mathbf{x};\gamma)$

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

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

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

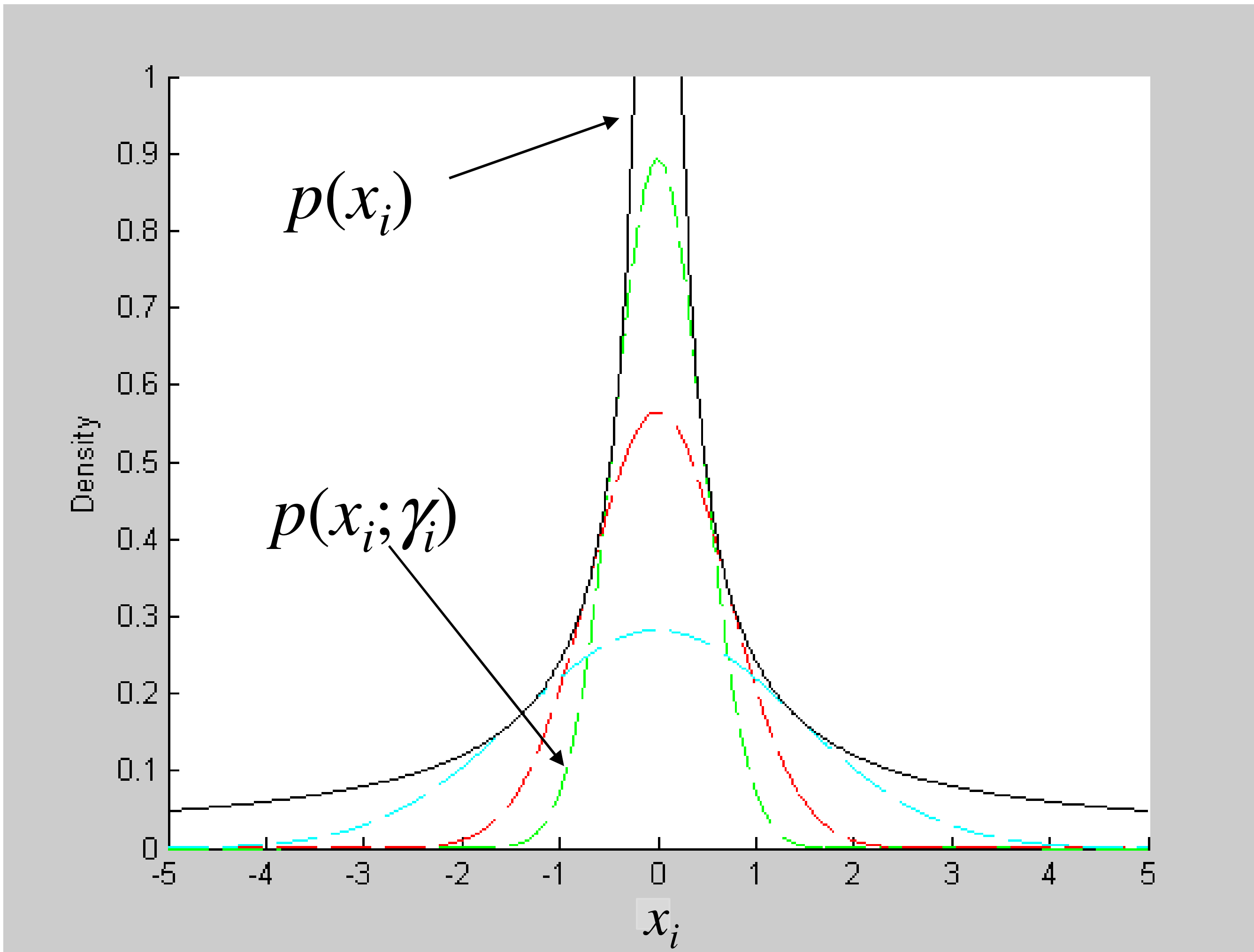
- Dropping the maximization gives the strict variational lower bound

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

- Yields a convenient class of scaled Gaussian approximations:

$$p\left(\mathbf{x};\boldsymbol{\gamma}\right) = \prod_i p\left(x_i;\gamma_i\right)$$

Example: Approximations to Sparse Prior



Step 4: Solving for the Optimal γ

- To find the best approximation, must solve

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

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

$$\begin{aligned} \hat{\boldsymbol{\gamma}} &= \arg\max_{\boldsymbol{\gamma}\geq 0} \int p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x};\boldsymbol{\gamma})\, d\mathbf{x} \\ &= \arg\min_{\boldsymbol{\gamma}\geq 0} \log\left|\Sigma_y\right| + \mathbf{y}^T \Sigma_y^{-1} \mathbf{y} - 2\sum_{i=1}^{m} \log \varphi\left(\gamma_i\right) \end{aligned}$$

where $\Sigma_y = \lambda I + \Phi\Gamma\Phi^T \qquad \Gamma = \mathrm{diag}\left(\boldsymbol{\gamma}\right)$

How difficult is finding the optimal parameters γ?

- 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

$$p(\mathbf{y} \mid \mathbf{x})\, p(\mathbf{x}; \hat{\boldsymbol{\gamma}}) = p(\mathbf{x}, \mathbf{y}; \hat{\boldsymbol{\gamma}}) \approx p(\mathbf{x}, \mathbf{y})$$

- By design, this approximation reflects significant mass in the full distribution $p(\mathbf{x},\mathbf{y})$.

- Also, it is easily normalized to form

$$p(\mathbf{x} \mid \mathbf{y}; \hat{\boldsymbol{\gamma}}) = N(\propto_x, \Sigma_x)$$

$$\propto_x \quad = \quad \mathrm{E}[\mathbf{x} \mid \mathbf{y}; \hat{\gamma}] \quad = \quad \hat{\Gamma}\Phi^T \left(\lambda I + \Phi\hat{\Gamma}\Phi^T\right)^{-1} \mathbf{y}$$

$$\Sigma_x \quad = \quad \mathrm{Cov}[\mathbf{x} \mid \mathbf{y}; \hat{\gamma}] \quad = \quad \hat{\Gamma} - \hat{\Gamma}\Phi^T \left(\lambda I + \Phi\hat{\Gamma}\Phi^T\right)^{-1} \Phi\hat{\Gamma}$$

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

$$p\left(\mathbf{x} \mid \mathbf{y}; \hat{\boldsymbol{\gamma}}\right) = N\left(\propto_x, \Sigma_x\right)$$

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

- Choose each additional row to minimize the differential entropy H:

$$H = \tfrac{1}{2}\log\left|\Sigma_x\right|, \quad \Sigma_x = \hat{\Gamma} - \hat{\Gamma}\Phi^T\left(\lambda I + \Phi\hat{\Gamma}\Phi^T\right)^{-1}\Phi\hat{\Gamma}$$

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 $\mathbf{x}$-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

$$\boldsymbol{\mu}_x = \mathrm{E}\left[\mathbf{x} \mid \mathbf{y}, \hat{\boldsymbol{\gamma}}\right] = \arg\min_{\mathbf{x}} \left\|\mathbf{y} - \Phi\mathbf{x}\right\|_2^2 + \lambda g_{\text{infer}}\left(\mathbf{x}\right)$$

with Bayesian inference penalty function

$$g_{\text{infer}}\left(\mathbf{x}\right) = \min_{\gamma \geq 0} \; \mathbf{x}^T \Gamma^{-1} \mathbf{x} + \log\left|\lambda I + \Phi\Gamma\Phi^T\right| - 2\sum_i \log\varphi\left(\gamma_i\right)$$

[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 $g_{\text{infer}}(\mathbf{x})$

- Penalty $g_{\text{infer}}(\mathbf{x})$ 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 ℓ_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 ℓ_2 update rules [Seeger, 2009; Wipf and Nagarajan, 2009].

- Algorithm can also be derived via EM [Tipping, 2001].

Reweighted ℓ_2 Implementation Example

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

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

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

- 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

$$\Sigma_x = \left[\lambda^{-1}\Phi^T\Phi + W^{(k)}\right]^{-1}, \quad W^{(k)} = \operatorname{diag}\left[\mathbf{w}^{(k)}\right]$$

Property II of Penalty $g_{\text{infer}}(\mathbf{x})$

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

1. $g_{\text{infer}}(\mathbf{x})$ is concave in $|x_i|$ for all i $\Rightarrow$ 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 ℓ_1 minimization

[Wipf, 2006; Wipf and Nagarajan, 2009]

Reweighted ℓ_1 Implementation Example

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

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

$$w_i^{(k+1)} \rightarrow \left[\phi_i^T\left(\sigma^2 I + \Phi\tilde{W}^{(k)}\,\mathrm{diag}\left[\mathbf{x}^{(k+1)}\right]\Phi^T\right)^{-1}\phi_i + a\right]^{\frac{1}{2}}$$

- 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.,

$$\left\|\mathbf{x}^{(k+1)}\right\|_0 \;\leq\; \left\|\mathbf{x}^{(k)}\right\|_0 \leq \left\|\mathbf{x}^{(\ell_1\text{-norm})}\right\|_0$$

Property III of Penalty $g_{\text{infer}}(\mathbf{x})$

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

- Consequently, in $\mathbf{x}$-space, the Bayesian Inference penalized regression problem

$$\min_{\mathbf{x}} \left\| \mathbf{y} - \Phi \mathbf{x} \right\|_2^2 + \lambda g_{\text{infer}}\left(\mathbf{x}\right)$$

  is generally smoother than associated MAP problem.

Student's t Example

- Assume the Student's t distribution for each unknown coefficient:

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

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 $\mathbf{y} = \Phi\mathbf{x}$ has a 1-D feasible region, i.e.,

$$m = n + 1$$

- Any feasible solution $\mathbf{x}$ will satisfy:

$$\mathbf{x} \;=\; \mathbf{x}_{\text{true}} + \alpha\mathbf{v}$$

where $\mathbf{v} \in \mathcal{N}\text{ull}(\Phi)$

α is a scalar

$\mathbf{x}_{\text{true}}$ is the true generative coefficients

- Can plot *penalty functions* vs. α to view local minima profile over the 1-D feasible region.

Empirical Example

- Generate an iid Gaussian random dictionary Φ with *10 rows* and *11 columns*.

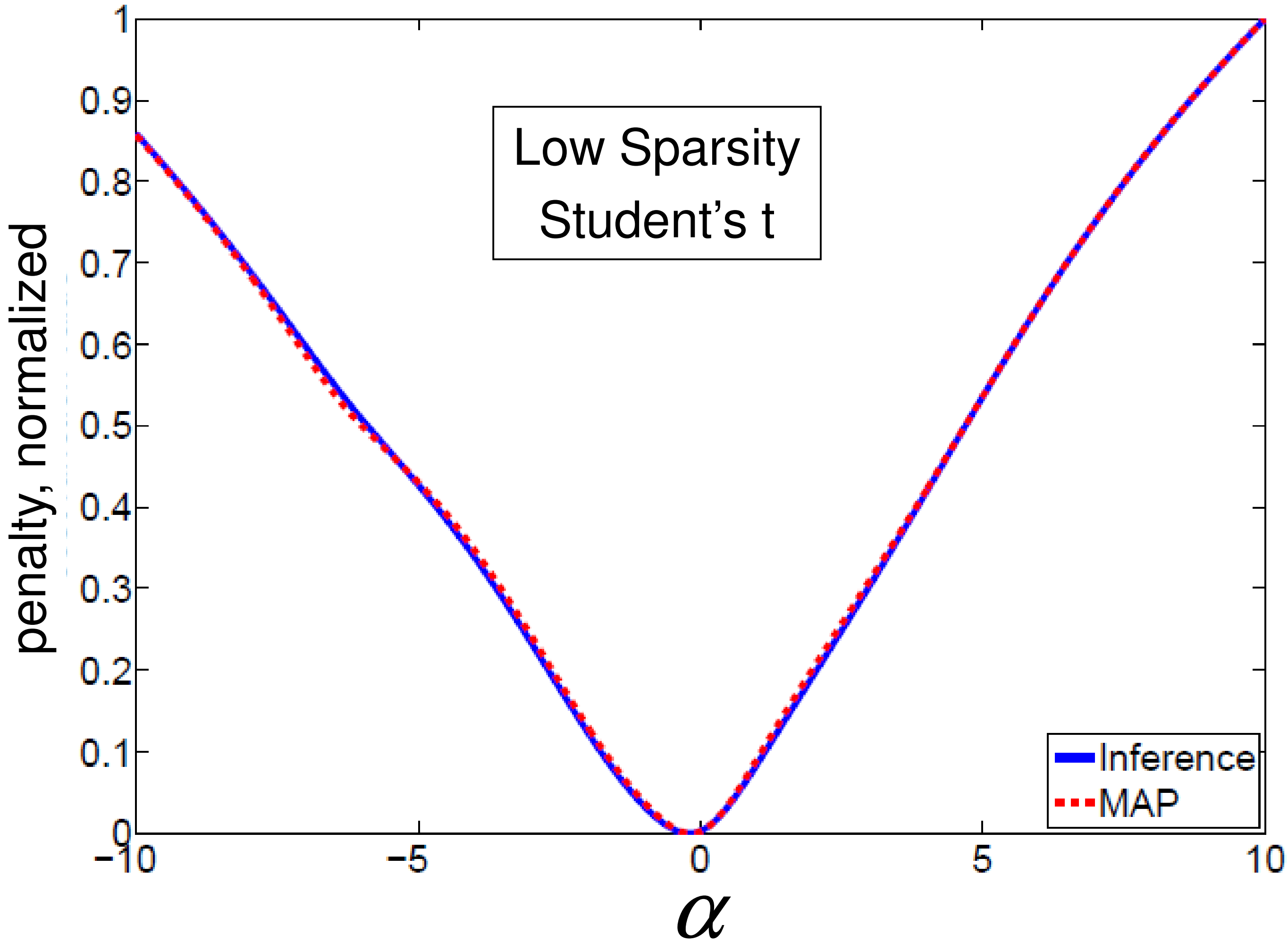
- Generate a sparse coefficient vector $\mathbf{x}_{\mathrm{true}}$ with *9 nonzeros* and Gaussian iid amplitudes.

- Compute signal $\mathbf{y} = \Phi \, \mathbf{x}_0$.

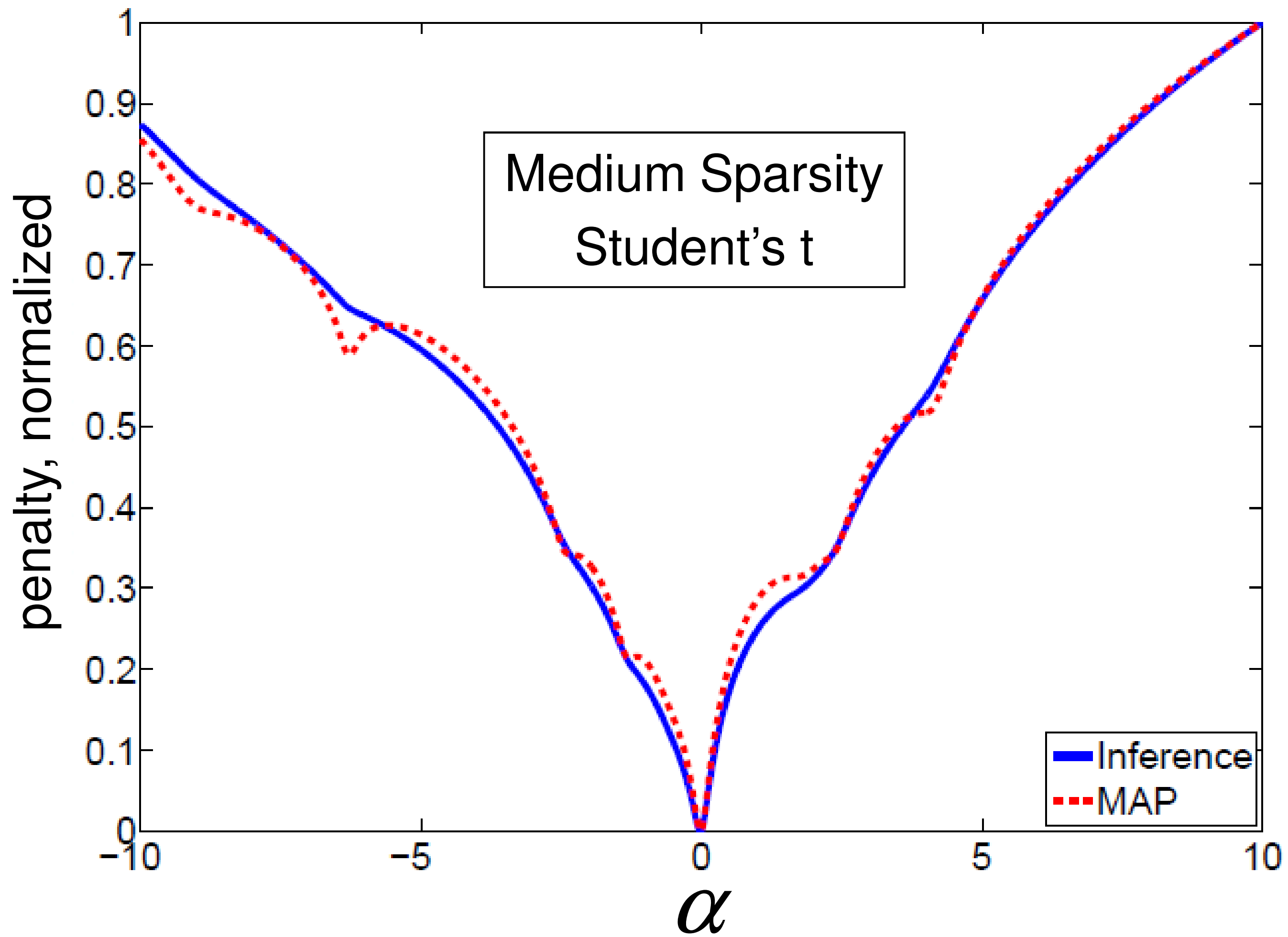
- Assume a Student's t prior on $\mathbf{x}$ with varying degrees of sparsity.

- Plot MAP/Bayesian inference penalties vs. α to compare local minima profiles over the 1-D feasible region.

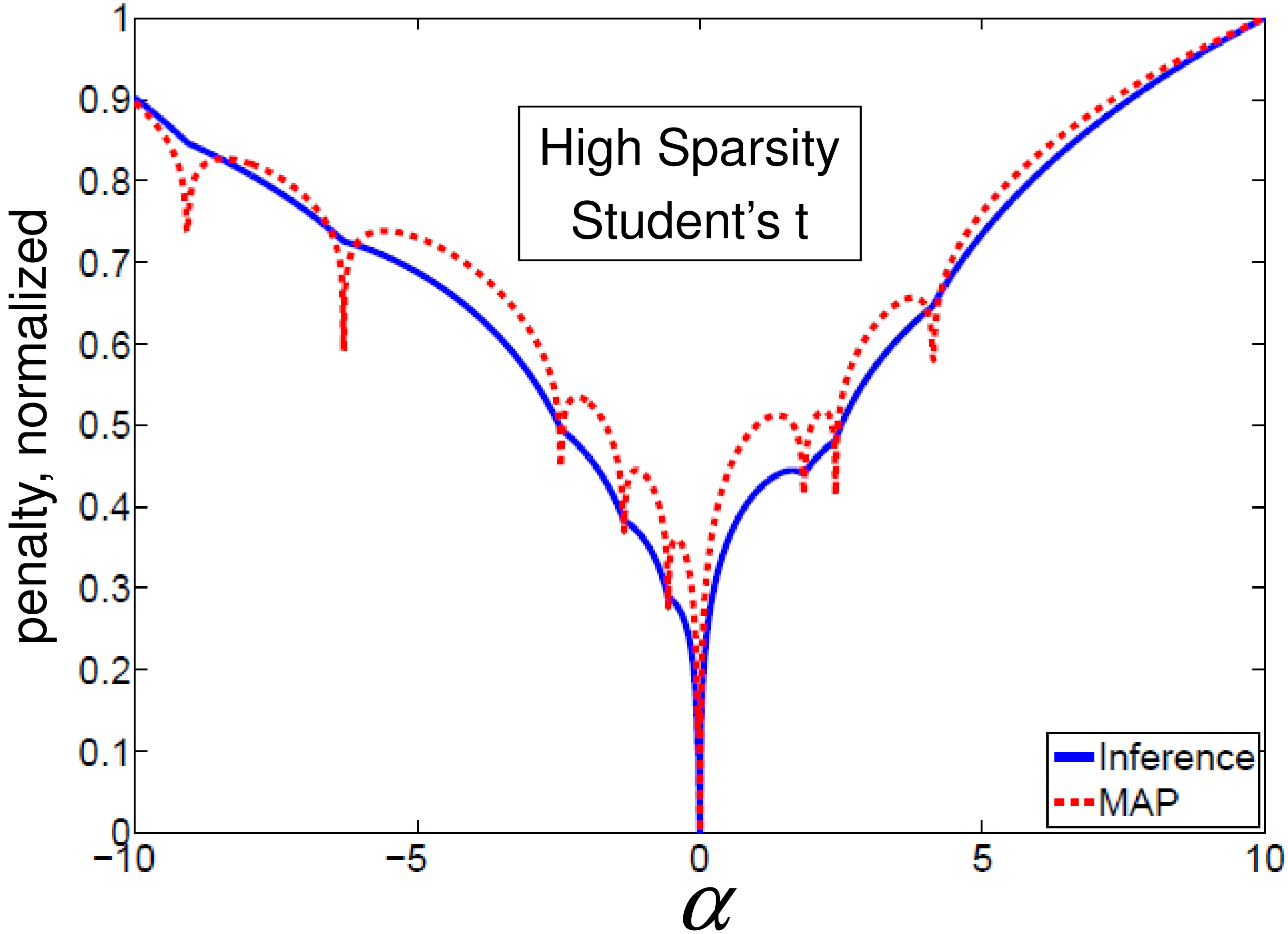
Local Minima Smoothing Example #1



Local Minima Smoothing Example #2



Local Minima Smoothing Example #3



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

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

$$\mathbf{x}_0 \triangleq \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \Phi\mathbf{x}$$

- **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):

$$\|\mathbf{x}\|_0 \leq n \quad \text{s.t.} \quad \mathbf{y} = \Phi\mathbf{x}$$

Visualization of Local Minima in ℓ_0 Norm

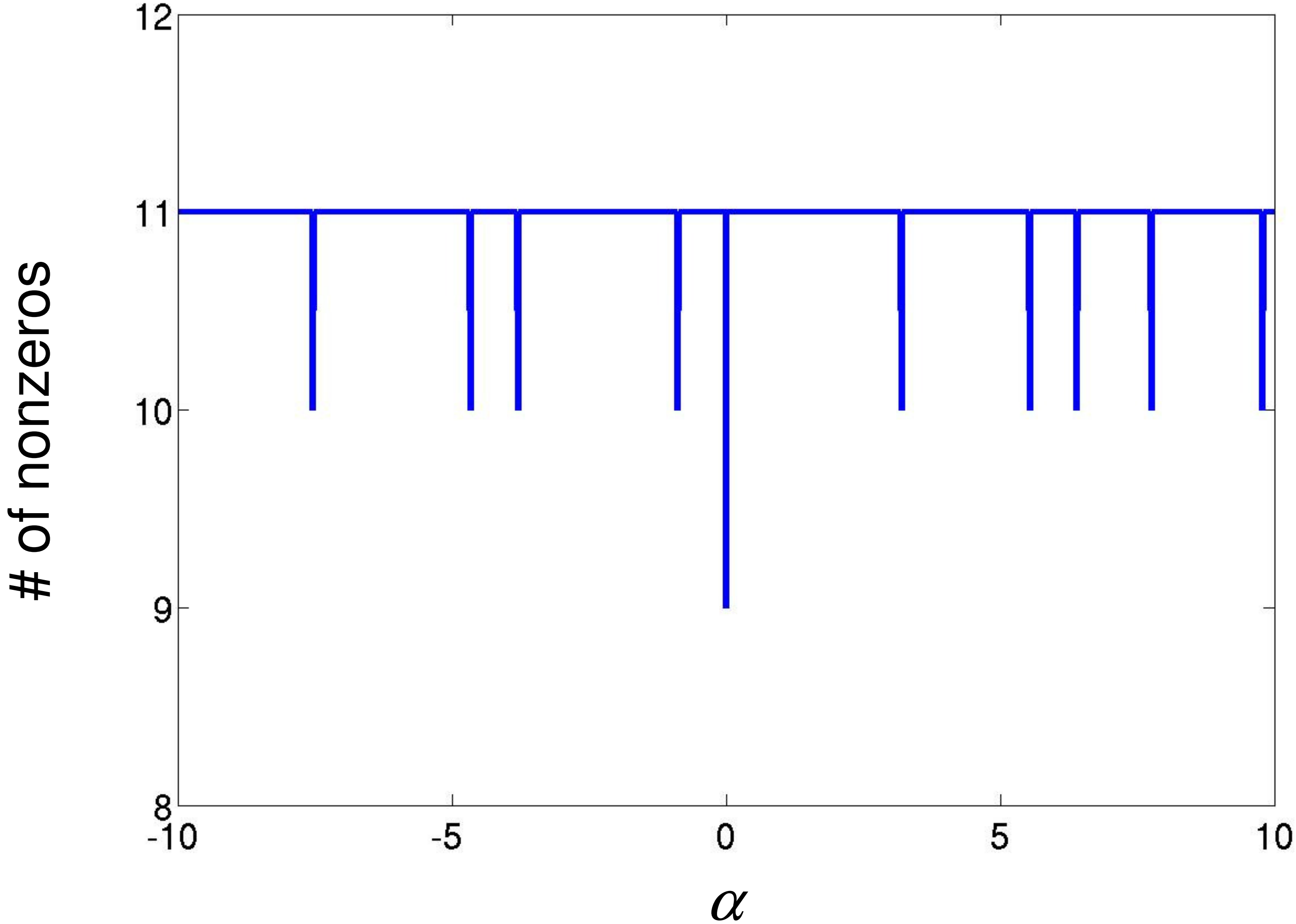
- Generate an iid Gaussian random dictionary Φ with *10 rows* and *11 columns*.

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

- Compute signal via $\mathbf{y} = \Phi\, \mathbf{x}_0$.

- Plot $\|\mathbf{x}\|_0$ vs. α (1-D null space dimension) to view local minima profile of the ℓ_0 norm over the 1-D feasible region.

ℓ_0 Norm in 1-D Feasible Region



Non-Separable Penalty Example

- Would like to smooth local minima while retaining same global solution as ℓ_0 at all times (unlike ℓ_1 norm)

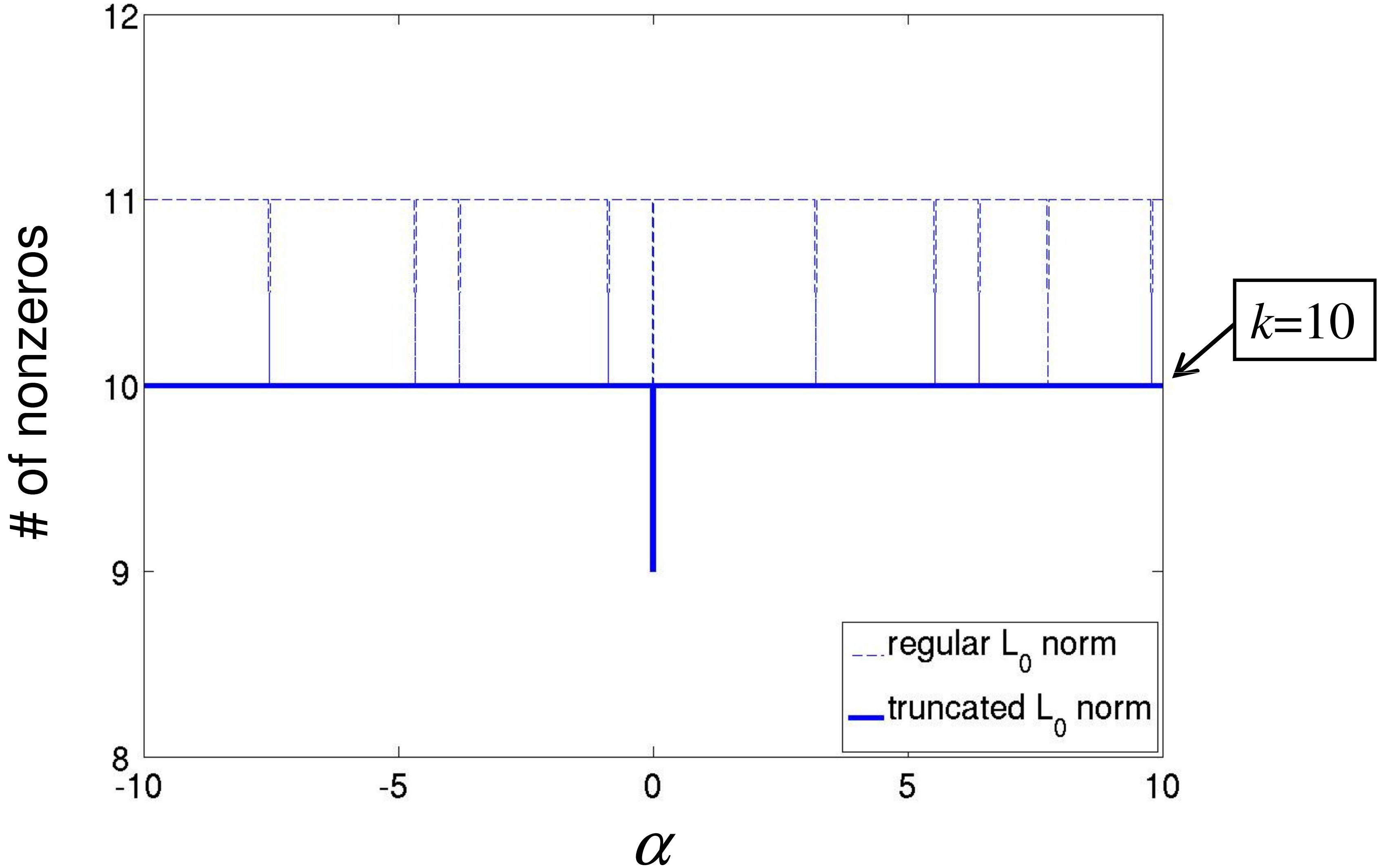
- This can be accomplished by a simple modification of the ℓ_0 penalty.

- **Truncated ℓ_0 penalty**:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\tilde{\mathbf{x}}\|_0 \quad \text{s.t.} \quad \begin{array}{l} \tilde{\mathbf{x}} = k \text{ largest elements of } \mathbf{x} \\ \mathbf{y} = \Phi \mathbf{x} \end{array}$$

- If $k < m$, then there will necessarily be fewer local minima; however, the implicit prior/penalty function is *non-separable*.

Truncated ℓ_0 Norm in 1-D Feasible Region



Using posterior mean estimator for finding maximally sparse solutions

Summary of why this could be a good idea:

1. If $-2\log\varphi(\gamma_i)$ 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 φ

- 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 $g_{\mathrm{SBL}}(\mathbf{x})$.

Advantages of Posterior Mean Estimator

Theorem

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

$$\arg \min_{\mathbf{x}:\, \mathbf{y}=\Phi\mathbf{x}} g_{\text{SBL}}(\mathbf{x}) = \arg \min_{\mathbf{x}:\, \mathbf{y}=\Phi\mathbf{x}} \|\mathbf{x}\|_0$$

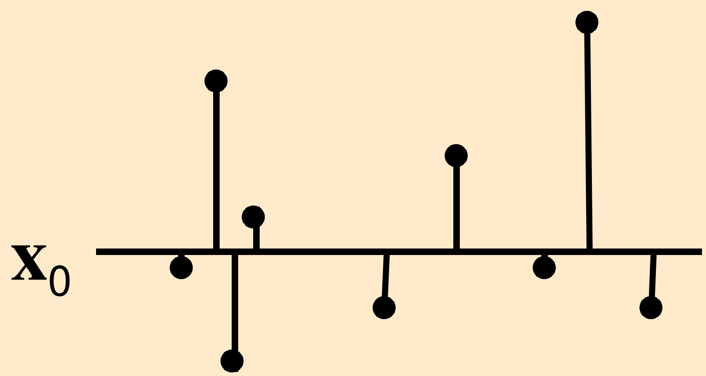
- No separable penalty $g(\mathbf{x}) = \sum_i g_i(x_i)$ satisfies this condition *and* has fewer minima than the SBL penalty in the feasible region.

[Wipf and Nagarajan, 2008]

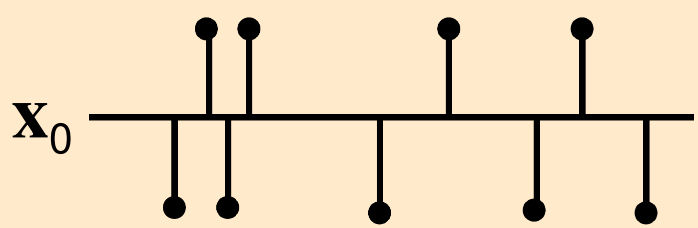
Conditions For a Single Minimum

Theorem

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



scaled coefficients (easy)



uniform coefficients (hard)

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

[Wipf and Nagarajan, 2009]

Empirical Example

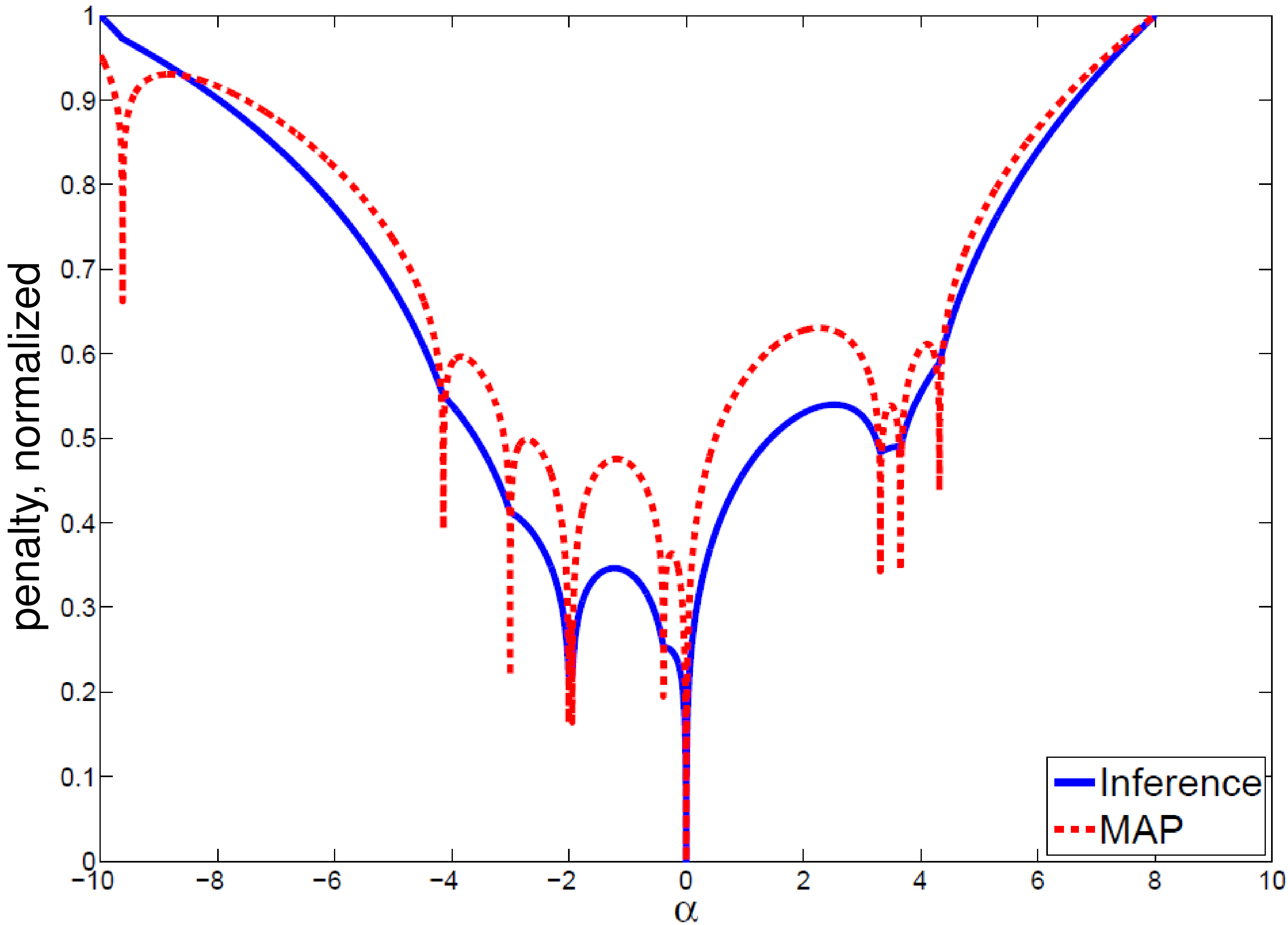
- Generate an iid Gaussian random dictionary Φ with *10 rows* and *11 columns*.

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

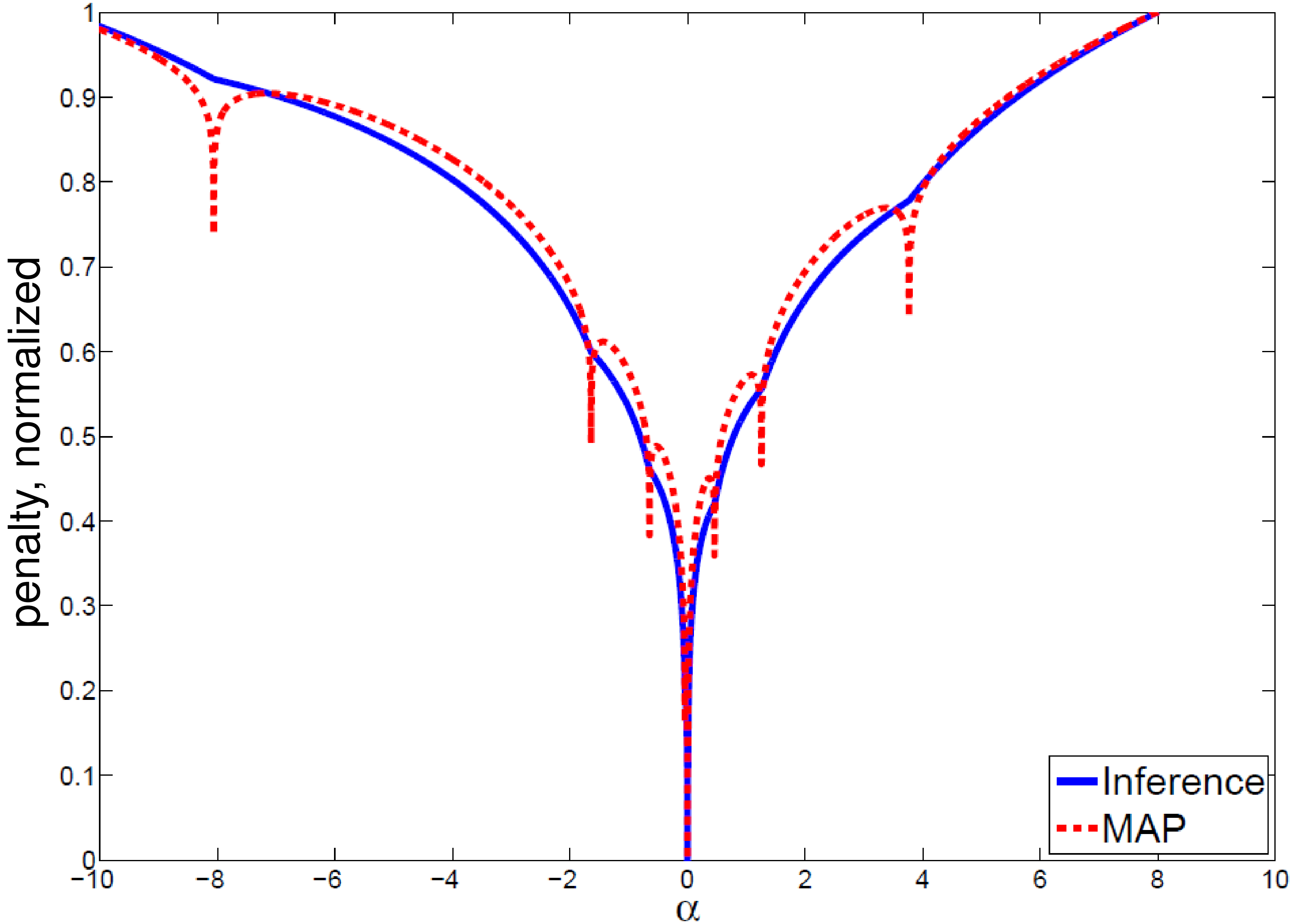
- Compute signal via $\mathbf{y} = \Phi\,\mathbf{x}_0$.

- Plot MAP/Bayesian inference penalty functions vs. α 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



Always Room for Improvement

Theorem

- Consider the noiseless sparse recovery problem.

$$\mathbf{x}_0 \triangleq \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \Phi\mathbf{x}$$

- Under very mild conditions, SBL with reweighted ℓ_1 implementation will:

1. Never do worse than the regular ℓ_1-norm solution
2. For any dictionary and sparsity profile, there will always be cases where it does better.

[Wipf and Nagarajan, 2010]

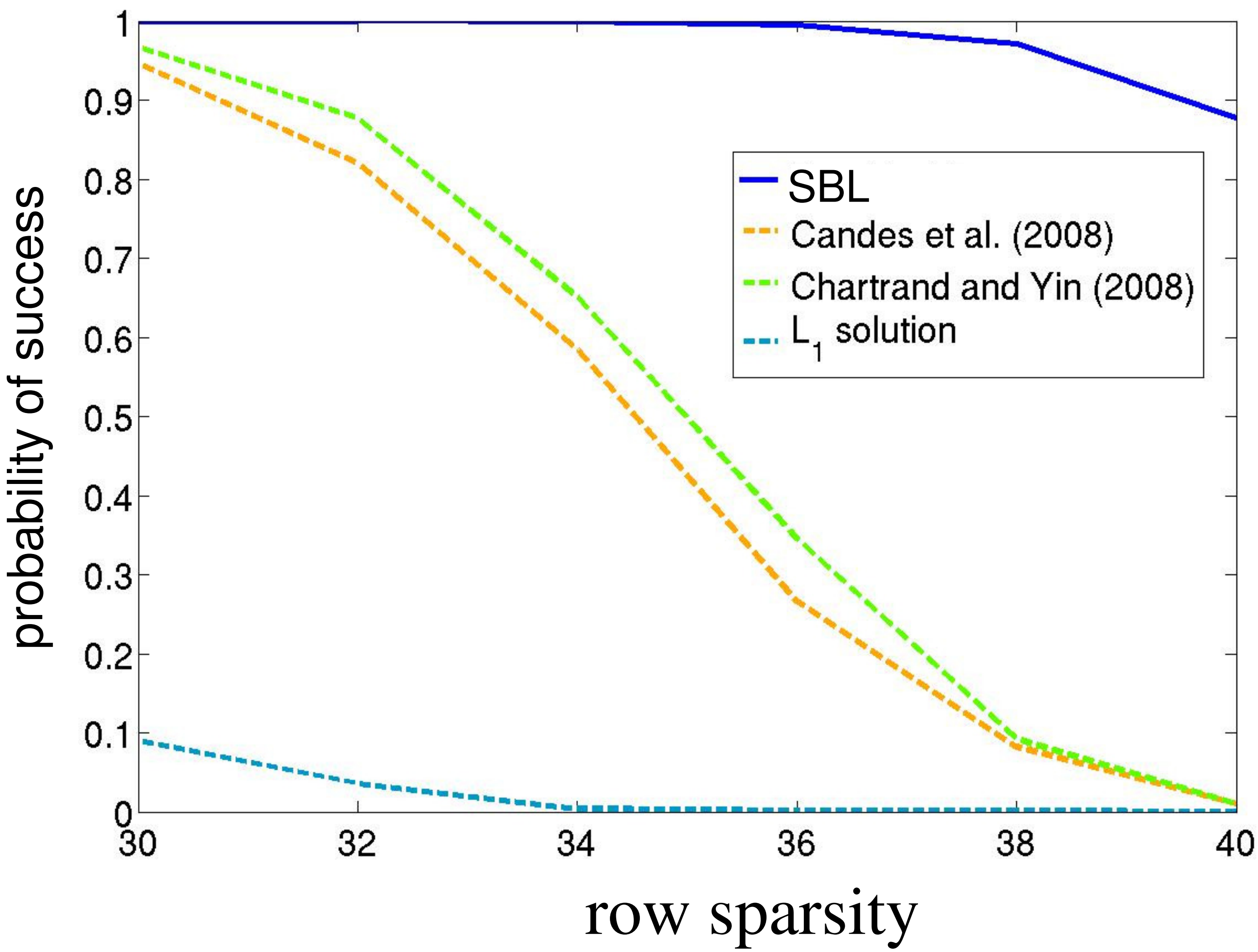
Empirical Example:
Simultaneous Sparse Approximation

- Generate data matrix via $\mathrm{Y} = \Phi \mathrm{X}_0$ (noiseless):
  - X_0 is 100-by-5 with random nonzero *rows*.
  - Φ is 50-by-100 with Gaussian iid entries

- Check if X_0 is recovered using various algorithms:

1. Generalized SBL , reweighted ℓ_2 implementation [Wipf and Nagarajan, 2010]
2. Candes et al. (2008) reweighted ℓ_1 method
3. Chartrand and Yin (2008) reweighted ℓ_2 method
4. ℓ_1 solution via Group Lasso [Yuan and Lin, 2006]

Empirical Results (1000 Trials)



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 ℓ_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

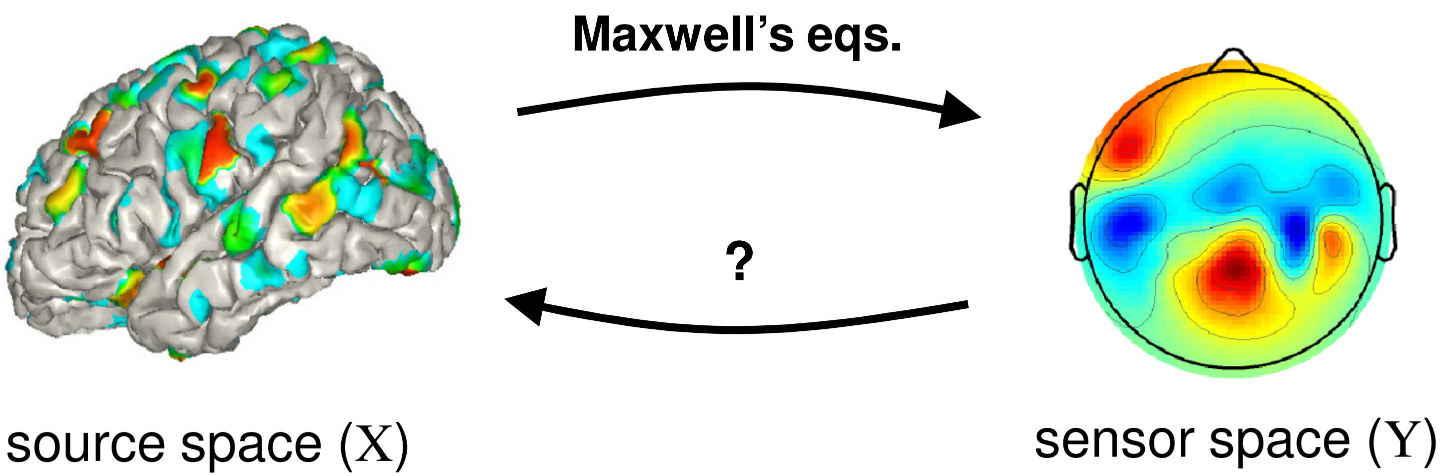
1. Recovering fiber track geometry from diffusion weighted MR images [Ramirez-Manzanares et al. 2007].

2. Multivariate autoregressive modeling of fMRI time series for functional connectivity analyses [Harrison et al. 2003].

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



The Dictionary Φ

- 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 Φ *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.

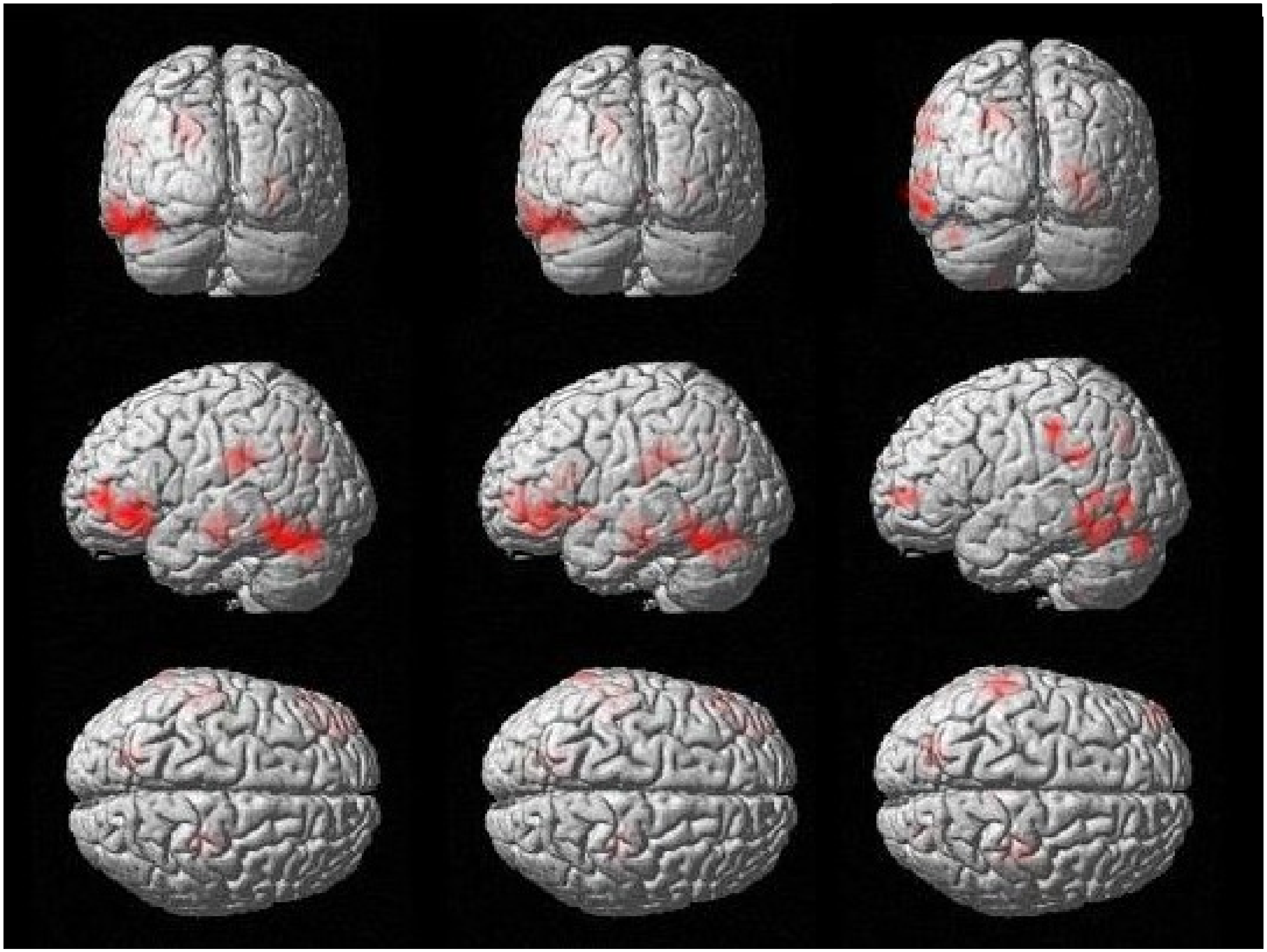
Empirical Results

1. Simulations with real brain noise/interference:
   - Generate damped sinusoidal sources
   - Map to sensors using Φ and apply *real brain noise, artifacts*

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

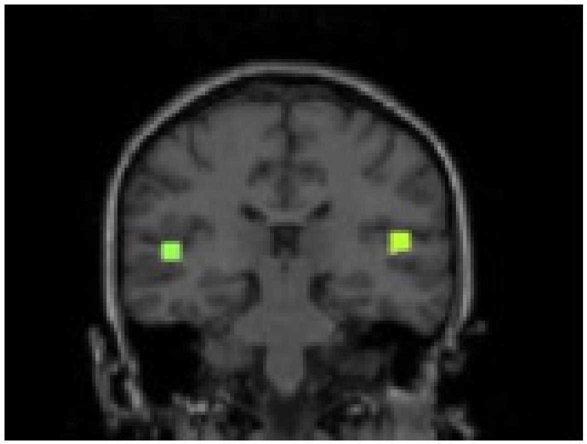
Compare localization results using MAP estimation and SBL posterior mean from Bayesian inference

MEG Source Reconstruction Example

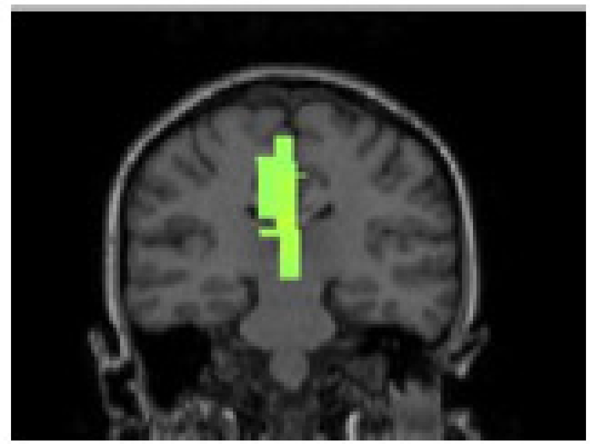


Ground Truth SBL Group Lasso

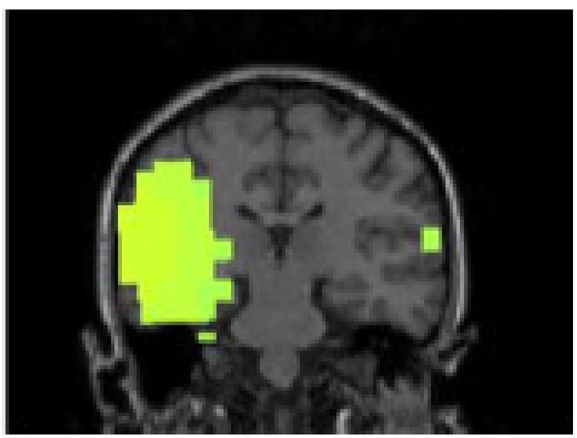
Real Data:
Auditory Evoked Field (AEF)



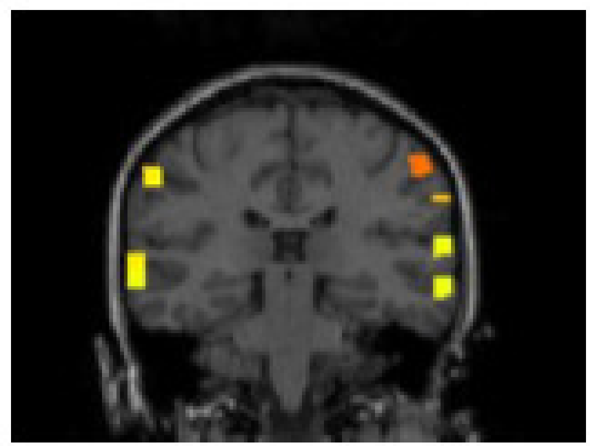
SBL



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.,

    $$n = 275 \text{ and } m = 100,000.$$

  - 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)
  - ℓ_1 norm minimization (Basis Pursuit, Lasso)
  - MAP methods (Reweighted ℓ_1 and ℓ_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