Gaussian Comparison Lemmas and Convex-Optimization

Babak Hassibi

joint work with
Samet Oymak, Christos Thrampoulidis and Ehsan Abbasi

California Institute of Technology

2016 London Workshop on Sparse Signal Processing
Imperial College, London, September 16, 2016

Outline

- **Introduction**
  - structured signal recovery
  - non-smooth convex optimization
  - LASSO and generalized LASSO; BPSK signal recovery
- **Comparison Lemmas**
  - Slepian, Gordon
- **Main Result**
  - squared error of generalized LASSO
  - Gaussian widths, statistical dimension
  - optimal parameter tuning
- **Generalizations**
  - other loss functions (Moreau envelopes)
  - other random matrix ensembles, universality
  - nonlinear measurements (one-bit compressed sensing)
- **Summmary and Conclusion**

Structured Signals

- We are increasingly confronted with very large data sets where we need to extract some *signal-of-interest*

Structured Signals

- We are increasingly confronted with very large data sets where we need to extract some *signal-of-interest*
  - machine learning, image processing, wireless comunications, signal processing, statistics, etc.

Structured Signals

- We are increasingly confronted with very large data sets where we need to extract some *signal-of-interest*
  - machine learning, image processing, wireless comunications, signal processing, statistics, etc.
  - sensor networks, social networks, massive MIMO, DNA microarrays, etc.

Structured Signals

- We are increasingly confronted with very large data sets where we need to extract some *signal-of-interest*
  - machine learning, image processing, wireless comunications, signal processing, statistics, etc.
  - sensor networks, social networks, massive MIMO, DNA microarrays, etc.
- On the face of it, this could lead to the *curse of dimensionality*

Structured Signals

- We are increasingly confronted with very large data sets where we need to extract some *signal-of-interest*
  - machine learning, image processing, wireless comunications, signal processing, statistics, etc.
  - sensor networks, social networks, massive MIMO, DNA microarrays, etc.
- On the face of it, this could lead to the *curse of dimensionality*
- Fortunately, in many applications, the signal of interest lives in a manifold of *much lower dimension* than that of the original ambient space

Structured Signals

- We are increasingly confronted with very large data sets where we need to extract some *signal-of-interest*
  - machine learning, image processing, wireless comunications, signal processing, statistics, etc.
  - sensor networks, social networks, massive MIMO, DNA microarrays, etc.
- On the face of it, this could lead to the *curse of dimensionality*
- Fortunately, in many applications, the signal of interest lives in a manifold of *much lower dimension* than that of the original ambient space
- In this setting, it is important to have signal recovery algorithms that are computationally efficient and that need not access the entire data directly (hence compressed recovery)

Non-Smooth Convex Optimization

- Non-smooth convex optimization has emerged as a tractable method to deal with structured signal recovery methods

Non-Smooth Convex Optimization

- Non-smooth convex optimization has emerged as a tractable method to deal with structured signal recovery methods
- Given the observations, $y \in \mathcal{R}^m$, we want to obtain some structured signal, $x \in \mathcal{R}^n$
  - a convex loss function $\mathcal{L}(x, y)$ (could be a log-likelihood function, e.g.)
  - a (non-smooth) convex *structure-inducing* regularizer $f(x)$

Non-Smooth Convex Optimization

- Non-smooth convex optimization has emerged as a tractable method to deal with structured signal recovery methods
- Given the observations, $y \in \mathcal{R}^m$, we want to obtain some structured signal, $x \in \mathcal{R}^n$
  - a convex loss function $\mathcal{L}(x, y)$ (could be a log-likelihood function, e.g.)
  - a (non-smooth) convex *structure-inducing* regularizer $f(x)$
- The generic problem is

$$\min_x \mathcal{L}(x,y) + \lambda f(x) \quad \text{or} \quad \min_{\mathcal{L}(x,y)\leq c_1} f(X) \quad \text{or} \quad \min_{f(x)\leq c_2} \mathcal{L}(x,y)$$

Non-Smooth Convex Optimization

$$\min_x \mathcal{L}(x,y) + \lambda f(x) \quad \text{or} \quad \min_{\mathcal{L}(x,y)\le c_1} f(X) \quad \text{or} \quad \min_{f(x)\le c_2} \mathcal{L}(x,y)$$

Non-Smooth Convex Optimization

$$\min_x \mathcal{L}(x,y) + \lambda f(x) \quad \text{or} \quad \min_{\mathcal{L}(x,y)\le c_1} f(X) \quad \text{or} \quad \min_{f(x)\le c_2} \mathcal{L}(x,y)$$

- **Algorithmic issues:**
  - scalable
  - distributed
  - etc.

Non-Smooth Convex Optimization

$$\min_{x} \mathcal{L}(x,y) + \lambda f(x) \quad \text{or} \quad \min_{\mathcal{L}(x,y)\leq c_1} f(X) \quad \text{or} \quad \min_{f(x)\leq c_2} \mathcal{L}(x,y)$$

- **Algorithmic issues:**
  - scalable
  - distributed
  - etc.
- **Analysis issues:**
  - can the *true* signal be recovered? (if so, when?)

Non-Smooth Convex Optimization

$$\min_{x} \mathcal{L}(x,y) + \lambda f(x) \quad \text{or} \quad \min_{\mathcal{L}(x,y)\leq c_1} f(X) \quad \text{or} \quad \min_{f(x)\leq c_2} \mathcal{L}(x,y)$$

- **Algorithmic issues:**
  - scalable
  - distributed
  - etc.
- **Analysis issues:**
  - can the *true* signal be recovered? (if so, when?)
  - if not, what is the quality of the recovered signal? (e.g., mean-square-error? probability of error?)

Non-Smooth Convex Optimization

$$\min_{x} \mathcal{L}(x,y) + \lambda f(x) \quad \text{or} \quad \min_{\mathcal{L}(x,y)\le c_1} f(X) \quad \text{or} \quad \min_{f(x)\le c_2} \mathcal{L}(x,y)$$

- **Algorithmic issues:**
  - scalable
  - distributed
  - etc.
- **Analysis issues:**
  - can the *true* signal be recovered? (if so, when?)
  - if not, what is the quality of the recovered signal? (e.g., mean-square-error? probability of error?)
  - how does the convex approach compare to one with no computational constraints?

Non-Smooth Convex Optimization

$$\min_x \mathcal{L}(x,y) + \lambda f(x) \quad \text{or} \quad \min_{\mathcal{L}(x,y)\le c_1} f(X) \quad \text{or} \quad \min_{f(x)\le c_2} \mathcal{L}(x,y)$$

- **Algorithmic issues:**
  - scalable
  - distributed
  - etc.
- **Analysis issues:**
  - can the *true* signal be recovered? (if so, when?)
  - if not, what is the quality of the recovered signal? (e.g., mean-square-error? probability of error?)
  - how does the convex approach compare to one with no computational constraints?
  - how to choose the regularizer $\lambda \ge 0$? (or the constraint bounds c_1 and c_2?)

Example: Noisy Compressed Sensing

Consider a "desired" signal $x \in \mathcal{R}^n$, which is k-sparse, i.e., has only $k < n$ (often $k \ll n$) non-zero entries. Suppose we make m noisy measurements of x using the $m \times n$ measurement matrix A to obtain

$$y = Ax + z.$$

Example: Noisy Compressed Sensing

Consider a "desired" signal $x \in \mathcal{R}^n$, which is k-sparse, i.e., has only $k < n$ (often $k \ll n$) non-zero entries. Suppose we make m noisy measurements of x using the $m \times n$ measurement matrix A to obtain

$$y = Ax + z.$$

How many measurements m do we need to find a good estimate of x?

Example: Noisy Compressed Sensing

Consider a "desired" signal $x \in \mathcal{R}^n$, which is k-sparse, i.e., has only $k < n$ (often $k \ll n$) non-zero entries. Suppose we make m noisy measurements of x using the $m \times n$ measurement matrix A to obtain

$$y = Ax + z.$$

How many measurements m do we need to find a good estimate of x? .

- Suppose each set of m columns of A are linearly independent. Then, if $m > k$, we can always find the *best k-sparse* solution to

$$\min_x \|y - Ax\|_2^2 \,,$$

via exhaustive search of $\left(\begin{array}{c} n \\ k \end{array} \right)$ *such least-squares problems*

Example: Noisy Compressed Sensing

Thus, the *information-theoretic* problem is perhaps not so challenging/interesting.

Example: Noisy Compressed Sensing

Thus, the *information-theoretic* problem is perhaps not so challenging/interesting. The *computational problem*, however, is:

Example: Noisy Compressed Sensing

Thus, the *information-theoretic* problem is perhaps not so challenging/interesting. The *computational problem*, however, is:

- Can we do this more efficiently? And for what values of m?

Example: Noisy Compressed Sensing

Thus, the *information-theoretic* problem is perhaps not so challenging/interesting. The *computational problem*, however, is:

- Can we do this more efficiently? And for what values of m?
- What about problems (such as low rank matrix recovery) where it is not possible to enumerate all structured signals?

LASSO

The LASSO algorithm was introduced by Tibshirani in 1996:

$$\hat{x} = \arg\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1,$$

where $\lambda \geq 0$ is a regularization parameter.

LASSO

The LASSO algorithm was introduced by Tibshirani in 1996:

$$\hat{x} = \arg\min_{x} \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1,$$

where $\lambda \geq 0$ is a regularization parameter.

Questions:

LASSO

The LASSO algorithm was introduced by Tibshirani in 1996:

$$\hat{x} = \arg\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1,$$

where $\lambda \geq 0$ is a regularization parameter.

Questions:

- How to choose λ?

LASSO

The LASSO algorithm was introduced by Tibshirani in 1996:

$$\hat{x} = \arg\min_{x} \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1,$$

where $\lambda \geq 0$ is a regularization parameter.

Questions:

- How to choose λ?
- What is the performance of the algorithm?

LASSO

The LASSO algorithm was introduced by Tibshirani in 1996:

$$\hat{x} = \arg\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda\|x\|_1,$$

where $\lambda \geq 0$ is a regularization parameter.

Questions:

- How to choose λ?
- What is the performance of the algorithm? For example, what is $E\|x - \hat{x}\|^2$?

Generalized LASSO

The generalized LASSO algorithm can be used to enforce other types of structures

$$\hat{x} = \arg\min_{x} \frac{1}{2} \|y - Ax\|_2^2 + \lambda f(x),$$

where $f(\cdot)$ is a *convex* regularizer.

Generalized LASSO

The generalized LASSO algorithm can be used to enforce other types of structures

$$\hat{x} = \arg\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda f(x),$$

where $f(\cdot)$ is a *convex* regularizer.

- $f(\cdot) = \|\cdot\|_1$ encourages sparsity

Generalized LASSO

The generalized LASSO algorithm can be used to enforce other types of structures

$$\hat{x} = \arg\min_{x} \frac{1}{2} \|y - Ax\|_2^2 + \lambda f(x),$$

where $f(\cdot)$ is a *convex* regularizer.

- $f(\cdot) = \|\cdot\|_1$ encourages sparsity
- $f(\cdot) = \|\cdot\|_\star$ encourages low rankness:

$$\hat{X} = \arg\min_{X} \frac{1}{2}\|y - A \cdot \mathrm{vec}(X)\|^2 + \lambda\|X\|_\star$$

Generalized LASSO

The generalized LASSO algorithm can be used to enforce other types of structures

$$\hat{x} = \arg\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda f(x),$$

where $f(\cdot)$ is a *convex* regularizer.

- $f(\cdot) = \|\cdot\|_1$ encourages sparsity
- $f(\cdot) = \|\cdot\|_\star$ encourages low rankness:

$$\hat{X} = \arg\min_X \frac{1}{2}\|y - A \cdot \mathrm{vec}(X)\|^2 + \lambda\|X\|_\star$$

- $f(\cdot) = \|\cdot\|_{1,2}$ (the mixed ℓ_1/ℓ_2 norm) encourages block-sparsity

$$\|x\|_{1,2} = \sum_b \|x_b\|_2.$$

Generalized LASSO

The generalized LASSO algorithm can be used to enforce other types of structures

$$\hat{x} = \arg\min_x \frac{1}{2}\|y - Ax\|_2^2 + \lambda f(x),$$

where $f(\cdot)$ is a *convex* regularizer.

- $f(\cdot) = \|\cdot\|_1$ encourages sparsity
- $f(\cdot) = \|\cdot\|_\star$ encourages low rankness:

$$\hat{X} = \arg\min_X \frac{1}{2}\|y - A \cdot \mathrm{vec}(X)\|^2 + \lambda\|X\|_\star$$

- $f(\cdot) = \|\cdot\|_{1,2}$ (the mixed ℓ_1/ℓ_2 norm) encourages block-sparsity

$$\|x\|_{1,2} = \sum_b \|x_b\|_2.$$

- $f(\cdot) = \|\cdot\|_\infty$ encourages constant-amplitude signals (BPSK, e.g.)

Generalized LASSO

The generalized LASSO algorithm can be used to enforce other types of structures

$$\hat{x} = \arg\min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda f(x),$$

where $f(\cdot)$ is a *convex* regularizer.

- $f(\cdot) = \|\cdot\|_1$ encourages sparsity
- $f(\cdot) = \|\cdot\|_\star$ encourages low rankness:

$$\hat{X} = \arg\min_X \frac{1}{2}\|y - A \cdot \mathrm{vec}(X)\|^2 + \lambda\|X\|_\star$$

- $f(\cdot) = \|\cdot\|_{1,2}$ (the mixed ℓ_1/ℓ_2 norm) encourages block-sparsity

$$\|x\|_{1,2} = \sum_b \|x_b\|_2.$$

- $f(\cdot) = \|\cdot\|_\infty$ encourages constant-amplitude signals (BPSK, e.g.)
- etc.

More General (Machine Learning) Problems

$$\min_x \mathcal{L}(x) + \lambda f(x),$$

where $\mathcal{L}(\cdot)$ is the so-called *loss function* and $f(\cdot)$ is the *regularizer*.

More General (Machine Learning) Problems

$$\min_x \mathcal{L}(x) + \lambda f(x),$$

where $\mathcal{L}(\cdot)$ is the so-called *loss function* and $f(\cdot)$ is the *regularizer*. For example,

- If the noise is Gaussian:

$$\hat{x} = \arg\min_x \|y - Ax\|_2 + \lambda f(x),$$

More General (Machine Learning) Problems

$$\min_x \mathcal{L}(x) + \lambda f(x),$$

where $\mathcal{L}(\cdot)$ is the so-called *loss function* and $f(\cdot)$ is the *regularizer*. For example,

- If the noise is Gaussian:

$$\hat{x} = \arg\min_x \|y - Ax\|_2 + \lambda f(x),$$

- If the noise is sparse:

$$\hat{x} = \arg\min_x \|y - Ax\|_1 + \lambda f(x),$$

More General (Machine Learning) Problems

$$\min_x \mathcal{L}(x) + \lambda f(x),$$

where $\mathcal{L}(\cdot)$ is the so-called *loss function* and $f(\cdot)$ is the *regularizer*.
For example,

- If the noise is Gaussian:

$$\hat{x} = \arg\min_x \|y - Ax\|_2 + \lambda f(x),$$

- If the noise is sparse:

$$\hat{x} = \arg\min_x \|y - Ax\|_1 + \lambda f(x),$$

- If the noise is bounded:

$$\hat{x} = \arg\min_x \|y - Ax\|_\infty + \lambda f(x),$$

The Squared Error of Generalized LASSO

$$\hat{x} = \arg\min_x \|y - Ax\|_2 + \lambda f(x)$$

- The LASSO algorithm has been extensively studied

The Squared Error of Generalized LASSO

$$\hat{x} = \arg\min_{x} \|y - Ax\|_2 + \lambda f(x)$$

- The LASSO algorithm has been extensively studied
- However, most performance bounds are rather loose

The Squared Error of Generalized LASSO

$$\hat{x} = \arg\min_{x} \|y - Ax\|_2 + \lambda f(x)$$

- The LASSO algorithm has been extensively studied
- However, most performance bounds are rather loose
- Can we compute $E\|x - \hat{x}\|^2$?

The Squared Error of Generalized LASSO

$$\hat{x} = \arg\min_{x} \|y - Ax\|_2 + \lambda f(x)$$

- The LASSO algorithm has been extensively studied
- However, most performance bounds are rather loose
- Can we compute $E\|x - \hat{x}\|^2$? Can we determine the optimal λ?

The Squared Error of Generalized LASSO

$$\hat{x} = \arg\min_x \|y - Ax\|_2 + \lambda f(x)$$

- The LASSO algorithm has been extensively studied
- However, most performance bounds are rather loose
- Can we compute $E\|x - \hat{x}\|^2$? Can we determine the optimal λ?

Turns out *we can*.....

Example

$\mathbf{X}_0 \in \mathbb{R}^{n\times n}$ is rank r. Observe, $\mathbf{y} = A \cdot \text{vec}(\mathbf{X}_0) + \mathbf{z}$, solve the Matrix LASSO,

$$\min_{\mathbf{X}} \{\|\mathbf{y} - A \cdot \text{vec}(\mathbf{X})\|_2 + \lambda \|\mathbf{X}\|_\star\}$$

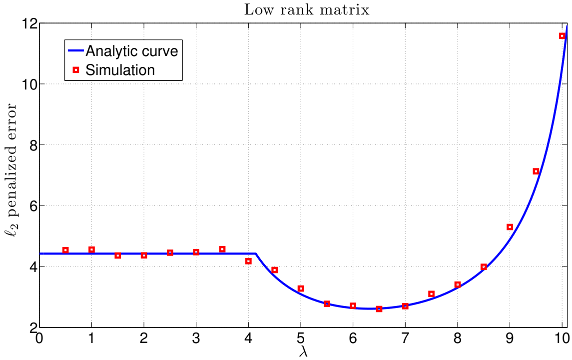


Figure: $n = 45$, $r = 6$, measurements $m = 0.6n^2$.

Recovering BPSK Signals

Consider

$$y = As + v,$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad , \quad s = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} \quad , \quad A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \quad , \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

Recovering BPSK Signals

Consider

$$y = As + v,$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad , \quad s = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} \quad , \quad A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \quad , \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

Asume BPSK signalling, i.e., $s_i \in \{\pm 1\}$.

Recovering BPSK Signals

Consider

$$y = As + v,$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad , \quad s = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} \quad , \quad A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \quad , \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

Asume BPSK signalling, i.e., $s_i \in \{\pm 1\}$. Furthermore, assume that A has iid $N(0,1)$ entries and that v has iid $N(0,\sigma^2)$ entries.

Recovering BPSK Signals

Consider

$$y = As + v,$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad , \quad s = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} \quad , \quad A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \quad , \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

Asume BPSK signalling, i.e., $s_i \in \{\pm 1\}$. Furthermore, assume that A has iid $N(0,1)$ entries and that v has iid $N(0,\sigma^2)$ entries. For a given SNR, $\sigma^2 = \frac{n}{\text{SNR}}$.

Recovering BPSK Signals

Consider

$$y = As + v,$$

where

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad , \quad s = \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} \quad , \quad A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \quad , \quad v = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

Asume BPSK signalling, i.e., $s_i \in \{\pm 1\}$. Furthermore, assume that A has iid $N(0,1)$ entries and that v has iid $N(0,\sigma^2)$ entries. For a given SNR, $\sigma^2 = \frac{n}{\mathrm{SNR}}$. The ML decoder is:

$$\hat{s} = \arg \min_{s_i \in \{\pm 1\}} \|y - As\|_2.$$

Box Relaxation

A natural convex relaxation is:

$$\hat{s} = \arg \min_{s_i \in [-1,1]} \|y - As\|_2.$$

Box Relaxation

A natural convex relaxation is:

$$\hat{s} = \arg \min_{s_i \in [-1,1]} \|y - As\|_2.$$

One can follow this by hard decision thresholding.

Box Relaxation

A natural convex relaxation is:

$$\hat{s} = \arg \min_{s_i \in [-1,1]} \|y - As\|_2.$$

One can follow this by hard decision thresholding.

This method is quite popular and referred to as *box relaxation*. But what is the BER?

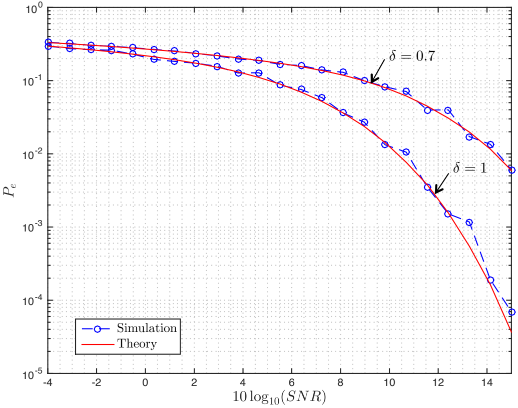


Figure: $n = 512$, $m = 358, 512$: Probability-of-error as a function of SNR

Where did this all come from....?

Slepian's Comparison Lemma (1962)

Slepian's Comparison Lemma (1962)

Let X_i and Y_i be two Gaussian processes with the same mean μ_i and variance σ_i^2, such that $\forall\ i, i'$

- $E(X_i - \mu_i)(X_{i'} - \mu_{i'}) \geq E(Y_i - \mu_i)(Y_{i'} - \mu_{i'})$

Then

Slepian's Comparison Lemma (1962)

Let X_i and Y_i be two Gaussian processes with the same mean μ_i and variance σ_i^2, such that $\forall\ i, i'$

- $E(X_i - \mu_i)(X_{i'} - \mu_{i'}) \geq E(Y_i - \mu_i)(Y_{i'} - \mu_{i'})$

Then

$$\text{Prob}\left(\max_i X_i \geq c\right) \overset{?}{\gtrless} \text{Prob}\left(\max_i Y_i \geq c\right)$$

Slepian's Comparison Lemma (1962)

Let X_i and Y_i be two Gaussian processes with the same mean μ_i and variance σ_i^2, such that $\forall\ i, i'$

- $E(X_i - \mu_i)(X_{i'} - \mu_{i'}) \geq E(Y_i - \mu_i)(Y_{i'} - \mu_{i'})$

Then

$$\text{Prob}\left(\max_i X_i \geq c\right) \leq \text{Prob}\left(\max_i Y_i \geq c\right)$$

Slepian's Comparison Lemma (1962)

- proof not too difficult, but not trivial, either
- lemma not generally true for non-Gaussian processes

Maximum Singular Value of a Gaussian Matrix

What is this good for?

Maximum Singular Value of a Gaussian Matrix

What is this good for?

Let $A \in \mathcal{R}^{m \times n}$ be a matrix with iid $N(0,1)$ entries and consider its maximum singular value:

$$\sigma_{\max}(A) = \|A\| = \max_{\|u\|=1} \max_{\|v\|=1} u^T A v.$$

Maximum Singular Value of a Gaussian Matrix

What is this good for?

Let $A \in \mathcal{R}^{m \times n}$ be a matrix with iid $N(0,1)$ entries and consider its maximum singular value:

$$\sigma_{\max}(A) = \|A\| = \max_{\|u\|=1} \max_{\|v\|=1} u^T A v.$$

Define the two Gaussian processes

$$X_{uv} = u^T A v + \gamma \quad \text{and} \quad Y_{uv} = u^T g + v^T h,$$

where $\gamma \in \mathcal{R}$, $g \in \mathcal{R}^m$ and $h \in \mathcal{R}^n$ have iid $N(0,1)$ entries.

Maximum Singular Value of a Gaussian Matrix

What is this good for?

Let $A \in \mathcal{R}^{m\times n}$ be a matrix with iid $N(0,1)$ entries and consider its maximum singular value:

$$\sigma_{\max}(A) = \|A\| = \max_{\|u\|=1} \max_{\|v\|=1} u^T A v.$$

Define the two Gaussian processes

$$X_{uv} = u^T A v + \gamma \quad \text{and} \quad Y_{uv} = u^T g + v^T h,$$

where $\gamma \in \mathcal{R}$, $g \in \mathcal{R}^m$ and $h \in \mathcal{R}^n$ have iid $N(0,1)$ entries. Then it is not hard to see that both processes have zero mean and variance 2.

Maximum Singular Value of a Gaussian Matrix

$$X_{uv} = u^T A v + \gamma \quad \text{and} \quad Y_{uv} = u^T g + v^T h,$$

Now,

$$EX_{uv}X_{u'v'} - EY_{uv}Y_{u'v'} = u^T u' v^T v' + 1 - u^T u' - v^T v' = (1 - u^T u')(1 - v^T v') \geq 0.$$

Maximum Singular Value of a Gaussian Matrix

$$X_{uv} = u^T A v + \gamma \quad \text{and} \quad Y_{uv} = u^T g + v^T h,$$

Now,

$$EX_{uv}X_{u'v'} - EY_{uv}Y_{u'v'} = u^T u' v^T v' + 1 - u^T u' - v^T v' = (1 - u^T u')(1 - v^T v') \geq 0.$$

Therefore from Slepian's lemma:

$$\underbrace{\text{Prob}\left(\max_{\|u\|=1}\max_{\|v\|=1} u^T A v + \gamma \geq c\right)}_{=\text{Prob}(\|A\|+\gamma\geq c)\geq\frac{1}{2}\text{Prob}(\|A\|\geq c)} \leq \underbrace{\text{Prob}\left(\max_{\|u\|=1}\max_{\|v\|=1} u^T g + v^T h \geq c\right)}_{\text{Prob}(\|g\|+\|h\|\geq c)}.$$

Maximum Singular Value of a Gaussian Matrix

$$X_{uv} = u^T A v + \gamma \quad \text{and} \quad Y_{uv} = u^T g + v^T h,$$

Now,

$$EX_{uv}X_{u'v'} - EY_{uv}Y_{u'v'} = u^T u' v^T v' + 1 - u^T u' - v^T v' = (1 - u^T u')(1 - v^T v') \geq 0.$$

Therefore from Slepian's lemma:

$$\underbrace{\text{Prob}\left(\max_{\|u\|=1}\max_{\|v\|=1} u^T A v + \gamma \geq c\right)}_{=\text{Prob}(\|A\|+\gamma\geq c)\geq\frac{1}{2}\text{Prob}(\|A\|\geq c)} \leq \underbrace{\text{Prob}\left(\max_{\|u\|=1}\max_{\|v\|=1} u^T g + v^T h \geq c\right)}_{\text{Prob}(\|g\|+\|h\|\geq c)}.$$

Since $\|g\| + \|h\|$ concentrates around $\sqrt{m} + \sqrt{n}$, this implies that the probability that $\|A\|$ (significantly) exceeds $\sqrt{m} + \sqrt{n}$ is very small.

Minimum Singular Value of a Gaussian Matrix

Let $A \in \mathcal{R}^{m\times n}$ $(m \leq n)$ be a matrix with iid $N(0,1)$ entries and consider its minimum singular value:

$$\sigma_{\min}(A) = \min_{\|u\|=1} \max_{\|v\|=1} u^T A v.$$

Minimum Singular Value of a Gaussian Matrix

Let $A \in \mathcal{R}^{m \times n}$ $(m \leq n)$ be a matrix with iid $N(0,1)$ entries and consider its minimum singular value:

$$\sigma_{\min}(A) = \min_{\|u\|=1} \max_{\|v\|=1} u^T A v.$$

Slepian's lemma does not apply.

Minimum Singular Value of a Gaussian Matrix

Let $A \in \mathcal{R}^{m \times n}$ $(m \leq n)$ be a matrix with iid $N(0,1)$ entries and consider its minimum singular value:

$$\sigma_{\min}(A) = \min_{\|u\|=1} \max_{\|v\|=1} u^T A v.$$

Slepian's lemma does not apply.

It took 24 years for there to be progress...

Gordon's Comparison Lemma (1988)

Let X_{ij} and Y_{ij} be two Gaussian processes with the same mean μ_{ij} and variance σ_{ij}^2, such that $\forall\ i, j, i', j'$

1. $E(X_{ij} - \mu_{ij})(X_{ij'} - \mu_{ij'}) \leq E(Y_{ij} - \mu_{ij})(Y_{ij'} - \mu_{ij'})$
2. $E(X_{ij} - \mu_{ij})(X_{i'j'} - \mu_{i'j'}) \geq E(Y_{ij} - \mu_{ij})(Y_{i'j'} - \mu_{i'j'})$

Then

$$\text{Prob}\left(\min_i \max_j X_{ij} \leq c\right) \overset{?}{\gtrless} \text{Prob}\left(\min_i \max_j Y_{ij} \leq c\right)$$

Gordon's Comparison Lemma (1988)

Let X_{ij} and Y_{ij} be two Gaussian processes with the same mean μ_{ij} and variance σ_{ij}^2, such that $\forall\ i, j, i', j'$

1. $E(X_{ij} - \mu_{ij})(X_{ij'} - \mu_{ij'}) \leq E(Y_{ij} - \mu_{ij})(Y_{ij'} - \mu_{ij'})$
2. $E(X_{ij} - \mu_{ij})(X_{i'j'} - \mu_{i'j'}) \geq E(Y_{ij} - \mu_{ij})(Y_{i'j'} - \mu_{i'j'})$

Then

$$\text{Prob}\left(\min_i \max_j X_{ij} \leq c\right) \leq \text{Prob}\left(\min_i \max_j Y_{ij} \leq c\right)$$

Gordon's Lemma (1988)

Let $G \in R^{m\times n}$, $\gamma \in R$, $g \in R^m$ and $h \in R^n$ have iid $N(0,1)$ entries, let S_x and S_y by compact sets, and $\psi(x,y)$ a continuous function.

Gordon's Lemma (1988)

Let $G \in R^{m \times n}$, $\gamma \in R$, $g \in R^m$ and $h \in R^n$ have iid $N(0,1)$ entries, let S_x and S_y by compact sets, and $\psi(x,y)$ a continuous function. Define:

$$\Phi(G,\gamma) = \min_{x \in S_x} \max_{y \in S_y} \; y^T G x + \gamma \|x\| \cdot \|y\| + \psi(x,y),$$

and

$$\phi(g,h) = \min_{x \in S_x} \max_{y \in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y).$$

Gordon's Lemma (1988)

Let $G \in R^{m\times n}$, $\gamma \in R$, $g \in R^m$ and $h \in R^n$ have iid $N(0,1)$ entries, let S_x and S_y by compact sets, and $\psi(x,y)$ a continuous function. Define:

$$\Phi(G,\gamma) = \min_{x\in S_x} \max_{y\in S_y} \; y^T Gx + \gamma\|x\|\cdot\|y\| + \psi(x,y),$$

and

$$\phi(g,h) = \min_{x\in S_x} \max_{y\in S_y} \; \|x\|g^T y + \|y\|h^T x + \psi(x,y).$$

Then it holds that:

$$\text{Prob}(\Phi(G,\gamma) \le c) \le \text{Prob}(\phi(g,h) \le c).$$

Gordon's Lemma (1988)

Let $G \in R^{m\times n}$, $\gamma \in R$, $g \in R^m$ and $h \in R^n$ have iid $N(0,1)$ entries, let S_x and S_y by compact sets, and $\psi(x,y)$ a continuous function. Define:

$$\Phi(G,\gamma) = \min_{x\in S_x} \max_{y\in S_y} \; y^T G x + \gamma\|x\|\cdot\|y\| + \psi(x,y),$$

and

$$\phi(g,h) = \min_{x\in S_x} \max_{y\in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y).$$

Then it holds that:

$$\mathrm{Prob}(\Phi(G,\gamma) \le c) \le \mathrm{Prob}(\phi(g,h) \le c).$$

- If c is a high probability lower bound on $\phi(\cdot,\cdot)$, same is true of $\Phi(\cdot,\cdot)$

Gordon's Lemma (1988)

Let $G \in R^{m \times n}$, $\gamma \in R$, $g \in R^m$ and $h \in R^n$ have iid $N(0,1)$ entries, let S_x and S_y by compact sets, and $\psi(x,y)$ a continuous function. Define:

$$\Phi(G,\gamma) = \min_{x \in S_x} \max_{y \in S_y} \; y^T G x + \gamma \|x\| \cdot \|y\| + \psi(x,y),$$

and

$$\phi(g,h) = \min_{x \in S_x} \max_{y \in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y).$$

Then it holds that:

$$\mathrm{Prob}(\Phi(G,\gamma) \le c) \le \mathrm{Prob}(\phi(g,h) \le c).$$

- If c is a high probability lower bound on $\phi(\cdot,\cdot)$, same is true of $\Phi(\cdot,\cdot)$
- Basis for "escape through mesh" and "Gaussian width"

Gordon's Lemma (1988)

Let $G \in R^{m \times n}$, $\gamma \in R$, $g \in R^m$ and $h \in R^n$ have iid $N(0,1)$ entries, let S_x and S_y by compact sets, and $\psi(x,y)$ a continuous function. Define:

$$\Phi(G,\gamma) = \min_{x \in S_x} \max_{y \in S_y} \; y^T G x + \gamma \|x\| \cdot \|y\| + \psi(x,y),$$

and

$$\phi(g,h) = \min_{x \in S_x} \max_{y \in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y).$$

Then it holds that:

$$\text{Prob}(\Phi(G,\gamma) \leq c) \leq \text{Prob}(\phi(g,h) \leq c).$$

- If c is a high probability lower bound on $\phi(\cdot,\cdot)$, same is true of $\Phi(\cdot,\cdot)$
- Basis for "escape through mesh" and "Gaussian width"
- Can be used to show that $\sigma_{\min}(A)$ behaves as $\sqrt{n} - \sqrt{m}$

A Stronger Version of Gordon's Lemma (TOH 2015)

$$\begin{cases} \Phi(G) & = & \min_{x \in S_x} \max_{y \in S_y} \; y^T G x + \psi(x,y) & \text{(PO)} \\ \phi(g,h) & = & \min_{x \in S_x} \max_{y \in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y) & \text{(AO)} \end{cases}$$

A Stronger Version of Gordon's Lemma (TOH 2015)

$$\begin{cases} \Phi(G) &= \min_{x\in S_x} \max_{y\in S_y} \; y^T Gx + \psi(x,y) & \text{(PO)} \\ \phi(g,h) &= \min_{x\in S_x} \max_{y\in S_y} \; \|x\|g^T y + \|y\|h^T x + \psi(x,y) & \text{(AO)} \end{cases}$$

Theorem

1. $Prob(\Phi(G) \leq c) \leq 2Prob(\phi(g,h) \leq c)$.

A Stronger Version of Gordon's Lemma (TOH 2015)

$$\begin{cases} \Phi(G) &= \min_{x \in S_x} \max_{y \in S_y} \; y^T G x + \psi(x,y) & \text{(PO)} \\ \phi(g,h) &= \min_{x \in S_x} \max_{y \in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y) & \text{(AO)} \end{cases}$$

Theorem

1. *Prob*$(\Phi(G) \le c) \le 2$*Prob*$(\phi(g,h) \le c)$.
2. *If* S_x *and* S_y *are* convex sets, *at least one of which is compact, and* $\psi(x,y)$ *is a convex-concave function, then*

A Stronger Version of Gordon's Lemma (TOH 2015)

$$\begin{cases} \Phi(G) = \min_{x \in S_x} \max_{y \in S_y} \; y^T G x + \psi(x,y) & \text{(PO)} \\ \phi(g,h) = \min_{x \in S_x} \max_{y \in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y) & \text{(AO)} \end{cases}$$

Theorem

1. *$Prob(\Phi(G) \leq c) \leq 2Prob(\phi(g,h) \leq c)$.*
2. *If S_x and S_y are* convex sets, *at least one of which is compact, and $\psi(x,y)$ is a convex-concave function, then*

$$Prob\left(|\Phi(G) - c| \geq \epsilon\right) \leq 2Prob\left(|\phi(g,h) - c| \geq \epsilon\right).$$

A Stronger Version of Gordon's Lemma (TOH 2015)

$$\begin{cases} \Phi(G) = \min_{x\in S_x} \max_{y\in S_y} \; y^T G x + \psi(x,y) & \text{(PO)} \\ \phi(g,h) = \min_{x\in S_x} \max_{y\in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y) & \text{(AO)} \end{cases}$$

Theorem

1. $Prob(\Phi(G) \le c) \le 2Prob(\phi(g,h) \le c)$.
2. *If S_x and S_y are* convex sets, *at least one of which is compact, and $\psi(x,y)$ is a convex-concave function, then*

$$Prob\left(|\Phi(G) - c| \ge \epsilon\right) \le 2Prob\left(|\phi(g,h) - c| \ge \epsilon\right).$$

3. *If, in addition, the optimization over x in (PO) is* strongly *convex,*

$$Prob(\hat{x}_\Phi \in S) \le 4Prob(\hat{x}_\phi \in S), \qquad \forall S.$$

A Stronger Version of Gordon's Lemma (TOH 2015)

$$\begin{cases} \Phi(G) = \min_{x\in S_x} \max_{y\in S_y} \; y^T G x + \psi(x,y) & \text{(PO)} \\ \phi(g,h) = \min_{x\in S_x} \max_{y\in S_y} \; \|x\| g^T y + \|y\| h^T x + \psi(x,y) & \text{(AO)} \end{cases}$$

Theorem

1. $Prob(\Phi(G) \leq c) \leq 2Prob(\phi(g,h) \leq c)$.
2. *If S_x and S_y are* convex sets, *at least one of which is compact, and $\psi(x,y)$ is a convex-concave function, then*

$$Prob\left(|\Phi(G) - c| \geq \epsilon\right) \leq 2Prob\left(|\phi(g,h) - c| \geq \epsilon\right).$$

3. *If, in addition, the optimization over x in (PO) is* strongly *convex,*

$$Prob(\hat{x}_\Phi \in S) \leq 4Prob(\hat{x}_\phi \in S), \quad \forall S.$$

4. *Under the above assumptions, $\hat{x}_\Phi$ and $\hat{x}_\phi$* asymptotically have the same empirical distribution.

Remarks

- In 3 take

$$S = \{x, |\|x\| - c| \geq \epsilon\}.$$

Remarks

- In 3 take

$$S = \{x, |\|x\| - c| \geq \epsilon\}.$$

Then 3 shows that if $\|\hat{x}_\phi\|$ concentrates to c, $\|\hat{x}_\Phi\|$ concentrates to the same value.

Remarks

- In 3 take
$$S = \{x, |\|x\| - c| \geq \epsilon\}.$$
Then 3 shows that if $\|\hat{x}_\phi\|$ concentrates to c, $\|\hat{x}_\Phi\|$ concentrates to the same value.
- 4 can be used to evaluate the probability-of-error of the PO by analyzing the AO.

Analysis of the BER for Box Relaxation

Wlog, assume that the all -1 vector was transmitted:

$$y = -A\mathbf{1} + v.$$

Analysis of the BER for Box Relaxation

Wlog, assume that the all -1 vector was transmitted:

$$y = -A1 + v.$$

Therefore

$$\min_{s_i \in [-1,1]} \|y - As\|_2 = \min_{s_i \in [-1,1]} \|v - A(\underbrace{s+1}_{t})\|_2 = \min_{t_i \in [0,2]} \|v - At\|_2.$$

Analysis of the BER for Box Relaxation

Wlog, assume that the all -1 vector was transmitted:

$$y = -A1 + v.$$

Therefore

$$\min_{s_i \in [-1,1]} \|y - As\|_2 = \min_{s_i \in [-1,1]} \|v - A(\underbrace{s+1}_{t})\|_2 = \min_{t_i \in [0,2]} \|v - At\|_2.$$

Note that BER $=$ Prob$(t_i \geq 1)$.

Analysis of the BER for Box Relaxation

Wlog, assume that the all -1 vector was transmitted:

$$y = -A\mathbf{1} + v.$$

Therefore

$$\min_{s_i \in [-1,1]} \|y - As\|_2 = \min_{s_i \in [-1,1]} \|v - A(\underbrace{s + 1}_{t})\|_2 = \min_{t_i \in [0,2]} \|v - At\|_2.$$

Note that $\text{BER} = \text{Prob}(t_i \geq 1)$. Writing this as a PO:

$$\min_{t_i \in [0,2]} \max_{\|u\|_2 \leq 1} u^T (v - At) = \min_{t_i \in [0,2]} \max_{\|u\|_2 \leq 1} u^T \begin{bmatrix} -A & \frac{v}{\sigma} \end{bmatrix} \begin{bmatrix} t \\ \sigma \end{bmatrix},$$

Analysis of the BER for Box Relaxation

Wlog, assume that the all -1 vector was transmitted:

$$y = -A\mathbf{1} + v.$$

Therefore

$$\min_{s_i \in [-1,1]} \|y - As\|_2 = \min_{s_i \in [-1,1]} \|v - A(\underbrace{s + 1}_{t})\|_2 = \min_{t_i \in [0,2]} \|v - At\|_2.$$

Note that BER $=$ Prob$(t_i \geq 1)$. Writing this as a PO:

$$\min_{t_i \in [0,2]} \max_{\|u\|_2 \leq 1} u^T(v - At) = \min_{t_i \in [0,2]} \max_{\|u\|_2 \leq 1} u^T \begin{bmatrix} -A & \frac{v}{\sigma} \end{bmatrix} \begin{bmatrix} t \\ \sigma \end{bmatrix},$$

the AO is

$$\min_{t_i \in [0,2]} \max_{\|u\|_2 \leq 1} \sqrt{\|t\|^2 + \sigma^2} u^T g + \|u\|(t^T h + \sigma\gamma).$$

Analysis of the AO

$$\min_{t_i \in [0,2]} \max_{\|u\|_2 \le 1} \sqrt{\|t\|^2 + \sigma^2} u^T g + \|u\|(t^T h + \sigma\gamma).$$

Analysis of the AO

$$\min_{t_i \in [0,2]} \max_{\|u\|_2 \leq 1} \sqrt{\|t\|^2 + \sigma^2} u^T g + \|u\|(t^T h + \sigma\gamma).$$

Optimizing over u is straightforward

$$\min_{t_i \in [0,2]} \sqrt{\|t\|^2 + \sigma^2} \underbrace{\|g\|}_{\approx\sqrt{m}} + t^T h.$$

Analysis of the AO

$$\min_{t_i \in [0,2]} \max_{\|u\|_2 \le 1} \sqrt{\|t\|^2 + \sigma^2} u^T g + \|u\|(t^T h + \sigma\gamma).$$

Optimizing over u is straightforward

$$\min_{t_i \in [0,2]} \sqrt{\|t\|^2 + \sigma^2} \underbrace{\|g\|}_{\approx \sqrt{m}} + t^T h.$$

Using $\sqrt{x} = \min_{\beta > 0} \frac{\beta x}{2} + \frac{1}{2\beta}$, we obtain

$$\min_{t_i \in [0,2], \beta > 0} \frac{\beta}{2}(\|t\|^2 + \sigma^2)m + \frac{1}{2\beta} + t^T h.$$

$$= \min_{\beta > 0} \quad \frac{\beta mn}{2\mathrm{SNR}} + \frac{1}{2\beta} + \sum_{i=1}^{n} \min_{t_i \in [0,2]} \left(\frac{\beta m t_i^2}{2} + h_i t_i \right).$$

Analysis of the AO

$$\min_{\beta>0} \quad \frac{\beta mn}{2\mathrm{SNR}} + \frac{1}{2\beta} + \sum_{i=1}^{n} \min_{t_i\in[0,2]} \left(\frac{\beta m t_i^2}{2} + h_i t_i \right).$$

Analysis of the AO

$$\min_{\beta>0} \quad \frac{\beta mn}{2\mathrm{SNR}} + \frac{1}{2\beta} + \sum_{i=1}^{n} \min_{t_i\in[0,2]} \left(\frac{\beta m t_i^2}{2} + h_i t_i \right).$$

The optimization over t is now separable and straightforward:

$$\min_{\beta>0} \quad \frac{\beta mn}{2\mathrm{SNR}} + \frac{1}{2\beta} + \sum_{i=1}^{n} \begin{cases} 0 & \text{if } h_i \geq 0 & (\hat{t}_i = 0) \\ -\frac{h_i^2}{2\beta m} & \text{if } -2\beta m \leq h_i \leq 0 & (\hat{t}_i = -\frac{h_i}{\beta m}) \\ 2\beta m + 2h_i & \text{if } h_i \leq -2\beta m & (\hat{t}_i = -2) \end{cases}$$

Analysis of the AO

$$\min_{\beta>0} \quad \frac{\beta mn}{2\mathrm{SNR}} + \frac{1}{2\beta} + \sum_{i=1}^{n} \min_{t_i\in[0,2]} \left(\frac{\beta m t_i^2}{2} + h_i t_i \right).$$

The optimization over t is now separable and straightforward:

$$\min_{\beta>0} \quad \frac{\beta mn}{2\mathrm{SNR}} + \frac{1}{2\beta} + \sum_{i=1}^{n} \begin{cases} 0 & \text{if } h_i \geq 0 & (\hat{t}_i = 0) \\ -\frac{h_i^2}{2\beta m} & \text{if } -2\beta m \leq h_i \leq 0 & (\hat{t}_i = -\frac{h_i}{\beta m}) \\ 2\beta m + 2h_i & \text{if } h_i \leq -2\beta m & (\hat{t}_i = -2) \end{cases}$$

The summation concentrates to:

$$\min_{\beta>0} \quad \frac{\beta mn}{2\mathrm{SNR}} + \frac{1}{2\beta} + n \left(-\int_{-2\beta m}^{0} \frac{h^2}{2\beta m} p(h)dh + \int_{-\infty}^{-2\beta m} (2\beta m + 2h) p(h)dh \right).$$

Analysis of the AO

Redefining βm to β, after some algebra, we get

$$\hat{\beta} = \arg\min_{\beta > 0} \quad \frac{\beta}{2\mathrm{SNR}} + \frac{1}{2\beta}\left(1 - \frac{n}{2m}\right) + \frac{n}{2\beta m}\int_{2\beta}^{\infty}(h - 2\beta)^2 p(h)dh.$$

Analysis of the AO

Redefining βm to β, after some algebra, we get

$$\hat{\beta} = \arg\min_{\beta>0} \quad \frac{\beta}{2\mathrm{SNR}} + \frac{1}{2\beta}\left(1 - \frac{n}{2m}\right) + \frac{n}{2\beta m}\int_{2\beta}^{\infty}(h-2\beta)^2 p(h)dh.$$

Recall

$$BER = \mathrm{Prob}(\hat{t}_i \geq 1) = \mathrm{Prob}(-\frac{h_i}{\hat{\beta}} \geq 1) = \mathrm{Prob}(-h_i \geq \hat{\beta}).$$

So that

$$\mathrm{BER} = \int_{\hat{\beta}}^{\infty} \frac{e^{-h^2/2}}{\sqrt{2\pi}} dh = Q(\hat{\beta}).$$

BER

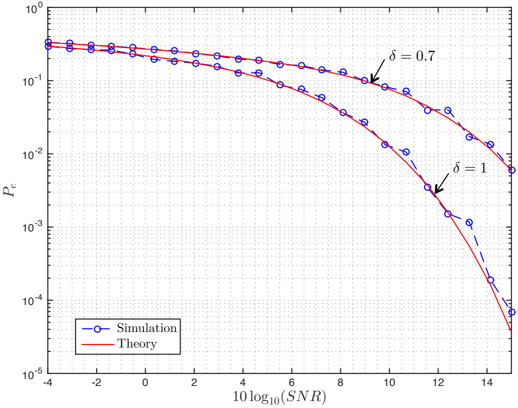


Figure: $n = 512$, $m = 358$: Probability-of-error as a function of SNR

Some Remarks

At high SNR, the value of $\hat{\beta}$ in the argument of the Q-function is large and therefore the intergral term in

$$\hat{\beta} = \arg\min_{\beta>0} \quad \frac{\beta}{2\mathrm{SNR}} + \frac{1}{2\beta}\left(1 - \frac{n}{2m}\right) + \frac{n}{2\beta m}\int_{2\beta}^{\infty}(h-2\beta)^2 p(h)dh.$$

can be ignored to obtain:

$$\hat{\beta} = \arg\min_{\beta>0} \quad \frac{\beta}{2\mathrm{SNR}} + \frac{1}{2\beta}\left(1 - \frac{n}{2m}\right).$$

Some Remarks

At high SNR, the value of $\hat{\beta}$ in the argument of the Q-function is large and therefore the intergral term in

$$\hat{\beta} = \arg\min_{\beta>0} \quad \frac{\beta}{2\mathrm{SNR}} + \frac{1}{2\beta}\left(1 - \frac{n}{2m}\right) + \frac{n}{2\beta m}\int_{2\beta}^{\infty}(h-2\beta)^2 p(h)dh.$$

can be ignored to obtain:

$$\hat{\beta} = \arg\min_{\beta>0} \quad \frac{\beta}{2\mathrm{SNR}} + \frac{1}{2\beta}\left(1 - \frac{n}{2m}\right).$$

This is a quadratic equation for $\hat{\beta}$ that can be straightforwardly solved to obtain:

$$\mathrm{BER} = Q\left(\sqrt{\left(\frac{m}{n} - \frac{1}{2}\right)\mathrm{SNR}}\right).$$

Some Remarks

The *matched filter bound* (MFB) assumes that all symbols $2, \ldots, n$ have been correctly decoded and looks at the probability of error of the first symbol.

Some Remarks

The *matched filter bound* (MFB) assumes that all symbols $2, \ldots, n$ have been correctly decoded and looks at the probability of error of the first symbol. It can be straightforwardly computed as

$$MFB = Q\left(\sqrt{\frac{m}{n}\text{SNR}}\right).$$

Some Remarks

The *matched filter bound* (MFB) assumes that all symbols $2, \ldots, n$ have been correctly decoded and looks at the probability of error of the first symbol. It can be straightforwardly computed as

$$MFB = Q\left(\sqrt{\frac{m}{n}\mathrm{SNR}}\right).$$

Thus, the box relaxation comes within $\log \frac{\frac{m}{n}}{\frac{m}{n}-\frac{1}{2}}$ db of the MFB.

Some Remarks

The *matched filter bound* (MFB) assumes that all symbols $2, \ldots, n$ have been correctly decoded and looks at the probability of error of the first symbol. It can be straightforwardly computed as

$$MFB = Q\left(\sqrt{\frac{m}{n}\mathrm{SNR}}\right).$$

Thus, the box relaxation comes within $\log \frac{\frac{m}{n}}{\frac{m}{n}-\frac{1}{2}}$ db of the MFB. For square systems ($m = n$) this is 3 db.

Some Remarks

The *matched filter bound* (MFB) assumes that all symbols $2, \ldots, n$ have been correctly decoded and looks at the probability of error of the first symbol. It can be straightforwardly computed as

$$MFB = Q\left(\sqrt{\frac{m}{n}\text{SNR}}\right).$$

Thus, the box relaxation comes within $\log \frac{\frac{m}{n}}{\frac{m}{n}-\frac{1}{2}}$ db of the MFB. For square systems ($m = n$) this is 3 db.

- In the AO, the events of making errors in each of the symbols were *independent*

Some Remarks

The *matched filter bound* (MFB) assumes that all symbols $2, \ldots, n$ have been correctly decoded and looks at the probability of error of the first symbol. It can be straightforwardly computed as

$$MFB = Q\left(\sqrt{\frac{m}{n}\mathrm{SNR}}\right).$$

Thus, the box relaxation comes within $\log \frac{\frac{m}{n}}{\frac{m}{n}-\frac{1}{2}}$ db of the MFB. For square systems ($m = n$) this is 3 db.

- In the AO, the events of making errors in each of the symbols were *independent*
- Therefor in the PO, for any fixed k symbols, the error events are also independent

Some Remarks

The *matched filter bound* (MFB) assumes that all symbols $2, \ldots, n$ have been correctly decoded and looks at the probability of error of the first symbol. It can be straightforwardly computed as

$$MFB = Q\left(\sqrt{\frac{m}{n}\text{SNR}}\right).$$

Thus, the box relaxation comes within $\log \frac{\frac{m}{n}}{\frac{m}{n}-\frac{1}{2}}$ db of the MFB. For square systems ($m = n$) this is 3 db.

- In the AO, the events of making errors in each of the symbols were *independent*
- Therefor in the PO, for any fixed k symbols, the error events are also independent
- This fact has far-reaching consequences for algorithms that can be applied to the output of the box relaxation

BER

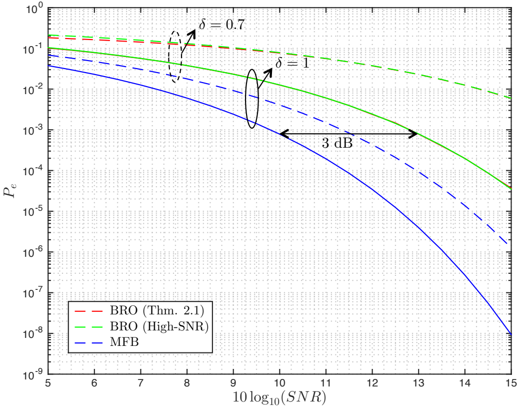


Figure: $n = 512$, $m = 358$: Probability-of-error as a function of SNR

Least-Squares

Suppose we are confronted with the *noisy* measurements:

$$y = Ax + z,$$

where $A \in \mathcal{R}^{m \times n}$ is the measurement matrix with iid $N(0,1)$ entries, $y \in \mathcal{R}^m$ is the measurement vector, $x_0 \in \mathcal{R}^n$ is the unknown desired signal, and $z \in \mathcal{R}^n$ is the unknown noise vector with iid $N(0,\sigma^2)$ entries.

Least-Squares

Suppose we are confronted with the *noisy* measurements:

$$y = Ax + z,$$

where $A \in \mathcal{R}^{m \times n}$ is the measurement matrix with iid $N(0,1)$ entries, $y \in \mathcal{R}^m$ is the measurement vector, $x_0 \in \mathcal{R}^n$ is the unknown desired signal, and $z \in \mathcal{R}^n$ is the unknown noise vector with iid $N(0,\sigma^2)$ entries. In the general case, to be meaningful, we require that

$$m \geq n.$$

Suppose we are confronted with the *noisy* measurements:

$$y = Ax + z,$$

where $A \in \mathcal{R}^{m\times n}$ is the measurement matrix with iid $N(0,1)$ entries, $y \in \mathcal{R}^m$ is the measurement vector, $x_0 \in \mathcal{R}^n$ is the unknown desired signal, and $z \in \mathcal{R}^n$ is the unknown noise vector with iid $N(0,\sigma^2)$ entries. In the general case, to be meaningful, we require that

$$m \geq n.$$

A popular method for recovering x, is the least-squares criterion

$$\min_x \|y - Ax\|_2.$$

Least-Squares

Suppose we are confronted with the *noisy* measurements:

$$y = Ax + z,$$

where $A \in \mathcal{R}^{m \times n}$ is the measurement matrix with iid $N(0,1)$ entries, $y \in \mathcal{R}^m$ is the measurement vector, $x_0 \in \mathcal{R}^n$ is the unknown desired signal, and $z \in \mathcal{R}^n$ is the unknown noise vector with iid $N(0,\sigma^2)$ entries. In the general case, to be meaningful, we require that

$$m \geq n.$$

A popular method for recovering x, is the least-squares criterion

$$\min_x \|y - Ax\|_2.$$

Let us analyze this using the stronger version of Gordon's lemma.

Least-Squares

To this end, define the estimation error $w = x_0 - x$, so that $y - Ax = Aw + z$.

Least-Squares

To this end, define the estimation error $w = x_0 - x$, so that $y - Ax = Aw + z$. Thus,

$$
\begin{aligned}
\min_x \|y - Ax\|_2 &= \min_w \|Aw + z\|_2 \\
&= \min_w \max_{\|u\| \le 1} u^T (Aw + z) = \min_w \max_{\|u\| \le 1} u^T \begin{bmatrix} A & \frac{1}{\sigma} z \end{bmatrix} \begin{bmatrix} w \\ \sigma \end{bmatrix}
\end{aligned}
$$

Least-Squares

To this end, define the estimation error $w = x_0 - x$, so that $y - Ax = Aw + z$. Thus,

$$\begin{aligned}\min_x \|y - Ax\|_2 &= \min_w \|Aw + z\|_2 \\ &= \min_w \max_{\|u\|\leq 1} u^T(Aw + z) = \min_w \max_{\|u\|\leq 1} u^T \begin{bmatrix} A & \frac{1}{\sigma}z \end{bmatrix} \begin{bmatrix} w \\ \sigma \end{bmatrix}\end{aligned}$$

This satisfies all the conditions of the lemma.

Least-Squares

To this end, define the estimation error $w = x_0 - x$, so that $y - Ax = Aw + z$. Thus,

$$\begin{aligned}\min_x \|y - Ax\|_2 &= \min_w \|Aw + z\|_2 \\ &= \min_w \max_{\|u\|\le 1} u^T(Aw+z) = \min_w \max_{\|u\|\le 1} u^T \left[\begin{array}{cc} A & \frac{1}{\sigma}z \end{array}\right] \left[\begin{array}{c} w \\ \sigma \end{array}\right]\end{aligned}$$

This satisfies all the conditions of the lemma. The simpler optimization is therefore:

$$\min_w \max_{\|u\|\le 1} \sqrt{\|w\|^2+\sigma^2}\, g^T u + \|u\| \left[\begin{array}{cc} h_w^T & h_\sigma \end{array}\right] \left[\begin{array}{c} w \\ \sigma \end{array}\right],$$

where $g = R^m$, $h_w = R^n$ and $h_\sigma \in R$ have iid $N(0,1)$ entries.

Least-Squares

$$\min_{w} \max_{\|u\|\leq 1} \sqrt{\|w\|^2+\sigma^2} g^T u + \|u\| \begin{bmatrix} h_w^T & h_\sigma \end{bmatrix} \begin{bmatrix} w \\ \sigma \end{bmatrix},$$

Least-Squares

$$\min_{w} \max_{\|u\|\leq 1} \sqrt{\|w\|^2+\sigma^2} g^T u + \|u\| \begin{bmatrix} h_w^T & h_\sigma \end{bmatrix} \begin{bmatrix} w \\ \sigma \end{bmatrix},$$

The maximization over u is straightforward:

$$\min_{w} \sqrt{\|w\|^2+\sigma^2}\|g\| + h_w^T w + h_\sigma \sigma.$$

Least-Squares

$$\min_{w}\max_{\|u\|\leq 1}\sqrt{\|w\|^2+\sigma^2}g^Tu+\|u\|\left[\begin{array}{cc} h_w^T & h_\sigma \end{array}\right]\left[\begin{array}{c} w \\ \sigma \end{array}\right],$$

The maximization over u is straightforward:

$$\min_{w}\sqrt{\|w\|^2+\sigma^2}\|g\|+h_w^Tw+h_\sigma\sigma.$$

Fixing the norm of $\|w\|=\alpha$, minimizing over the direction of w is straightforward:

$$\min_{\alpha\geq 0}=\sqrt{\alpha^2+\sigma^2}\|g\|-\alpha\|h_w\|+h_\sigma\sigma.$$

Least-Squares

$$\min_{w} \max_{\|u\|\leq 1} \sqrt{\|w\|^2+\sigma^2} g^T u + \|u\| \left[\begin{array}{cc} h_w^T & h_\sigma \end{array} \right] \left[\begin{array}{c} w \\ \sigma \end{array} \right],$$

The maximization over u is straightforward:

$$\min_{w} \sqrt{\|w\|^2+\sigma^2}\|g\| + h_w^T w + h_\sigma \sigma.$$

Fixing the norm of $\|w\| = \alpha$, minimizing over the direction of w is straightforward:

$$\min_{\alpha\geq 0} = \sqrt{\alpha^2+\sigma^2}\|g\| - \alpha\|h_w\| + h_\sigma \sigma.$$

Differentiating over α gives the solution:

$$\frac{\alpha^2}{\sigma^2} = \frac{\|h_w\|^2}{\|g\|^2-\|h_w\|^2} \rightarrow \frac{n}{m-n}.$$

Least-Squares

Thus, in summary:

$$\frac{E\|\hat{x} - x_0\|^2}{\sigma^2} \to \frac{n}{m-n}.$$

Least-Squares

Thus, in summary:

$$\frac{E\|\hat{x}-x_0\|^2}{\sigma^2} \to \frac{n}{m-n}.$$

Of course, in the least-squares case, we need not use all this machinery since the solutions are famously given by:

$$\hat{x} = \left(A^T A\right)^{-1} A^T y \quad \text{and} \quad E\|x_0 - \hat{x}\|_2^2 = \sigma^2 \text{trace}\left(A^T A\right)^{-1}.$$

Least-Squares

Thus, in summary:

$$\frac{E\|\hat{x} - x_0\|^2}{\sigma^2} \to \frac{n}{m-n}.$$

Of course, in the least-squares case, we need not use all this machinery since the solutions are famously given by:

$$\hat{x} = \left(A^T A\right)^{-1} A^T y \quad \text{and} \quad E\|x_0 - \hat{x}\|_2^2 = \sigma^2 \text{trace} \left(A^T A\right)^{-1}.$$

When A has iid $N(0,1)$ entries, $A^T A$ is a *Wishart matrix* whose asymptotic eigendistribution is well known, from which we obtain

$$\frac{E\|x - \hat{x}\|_2^2}{\sigma^2} \to \frac{n}{m-n}.$$

Back to the Squared Error of Generalized LASSO

However, for generalized LASSO, we do not have closed form solutions and the machinery becomes very useful:

$$\hat{x} = \arg\min_{x} \|y - Ax\|_2 + \lambda f(x)$$

Back to the Squared Error of Generalized LASSO

However, for generalized LASSO, we do not have closed form solutions and the machinery becomes very useful:

$$\hat{x} = \arg\min_{x} \|y - Ax\|_2 + \lambda f(x)$$

Using the same argument as before, we obtain the (AO):

$$\min_{w} \max_{\|u\| \leq 1} \sqrt{\|w\|^2 + \sigma^2} g^T u + \|u\| \begin{bmatrix} h_w^T & h_\sigma \end{bmatrix} \begin{bmatrix} w \\ \sigma \end{bmatrix} + \lambda f(x_0 - w).$$

Back to the Squared Error of Generalized LASSO

However, for generalized LASSO, we do not have closed form solutions and the machinery becomes very useful:

$$\hat{x} = \arg\min_{x} \|y - Ax\|_2 + \lambda f(x)$$

Using the same argument as before, we obtain the (AO):

$$\min_{w} \max_{\|u\|\leq 1} \sqrt{\|w\|^2 + \sigma^2} g^T u + \|u\| \begin{bmatrix} h_w^T & h_\sigma \end{bmatrix} \begin{bmatrix} w \\ \sigma \end{bmatrix} + \lambda f(x_0 - w).$$

Or:

$$\min_{w} \sqrt{\|w\|^2 + \sigma^2}\|g\| + h_w^T w + h_\sigma \sigma + \lambda f(x_0 - w).$$

Squared Error of Generalized LASSO $\sigma \to 0$

$$\min_{w} \sqrt{\|w\|^2 + \sigma^2}\|g\| + h_w^T w + h_\sigma \sigma + \lambda f(x_0 - w).$$

While this can be analyzed in full generality, it is instructive to focus on the low noise, $\sigma \to 0$, case.

Squared Error of Generalized LASSO $\sigma \to 0$

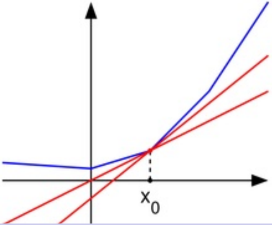
$$\min_{w} \sqrt{\|w\|^2+\sigma^2}\|g\| + h_w^T w + h_\sigma \sigma + \lambda f(x_0 - w).$$

While this can be analyzed in full generality, it is instructive to focus on the low noise, $\sigma \to 0$, case. Here $\|w\|$ will be small and we may therefore write

$$f(x_0 - w) \gtrsim f(x_0) + \sup_{s \in \partial f(\mathbf{x}_0)} s^T(-w),$$

where $\partial f(\mathbf{x}_0)$ is the subgradient of $f(\cdot)$ evaluated at x_0, and defined as

$$\partial f(\mathbf{x}_0) = \left\{ s | f(x + x_0) \geq f(x_0) + s^T x, \forall x \right\}.$$



Subgradients

The subgradient of a *convex function* is a *convex set*.

Subgradients

The subgradient of a *convex function* is a *convex set*. In most cases of interest subgradients are easy to compute. Here are some examples:

Subgradients

The subgradient of a *convex function* is a *convex set*. In most cases of interest subgradients are easy to compute. Here are some examples:

- $f(x) = \|x\|_1$ and $x_0 = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$:

Subgradients

The subgradient of a *convex function* is a *convex set*. In most cases of interest subgradients are easy to compute. Here are some examples:

- $f(x) = \|x\|_1$ and $x_0 = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$:

$$\partial f(\mathbf{x}_0) = \left\{ \begin{bmatrix} \mathrm{sign}(\xi) \\ s \end{bmatrix}, \|s\|_\infty \leq 1 \right\}.$$

Subgradients

The subgradient of a *convex function* is a *convex set*. In most cases of interest subgradients are easy to compute. Here are some examples:

- $f(x) = \|x\|_1$ and $x_0 = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$:

$$\partial f(\mathbf{x}_0) = \left\{ \begin{bmatrix} \text{sign}(\xi) \\ s \end{bmatrix}, \|s\|_\infty \leq 1 \right\}.$$

- $f(X) = \|X\|_\star$ and $X_0 = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^*$:

Subgradients

The subgradient of a *convex function* is a *convex set*. In most cases of interest subgradients are easy to compute. Here are some examples:

- $f(x) = \|x\|_1$ and $x_0 = \left[\begin{array}{c} \xi \\ 0 \end{array}\right]$:

$$\partial f(\mathbf{x}_0) = \left\{ \left[\begin{array}{c} \text{sign}(\xi) \\ s \end{array}\right], \|s\|_\infty \leq 1 \right\}.$$

- $f(X) = \|X\|_\star$ and $X_0 = U \left[\begin{array}{cc} \Sigma & 0 \\ 0 & 0 \end{array}\right] V^*$:

$$\partial f(\mathbf{x}_0) = \left\{ U \left[\begin{array}{cc} I & 0 \\ 0 & D \end{array}\right] V^*, |d_i| \leq 1 \right\}.$$

Subgradients

The subgradient of a *convex function* is a *convex set*. In most cases of interest subgradients are easy to compute. Here are some examples:

- $f(x) = \|x\|_1$ and $x_0 = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$:

$$\partial f(\mathbf{x}_0) = \left\{ \begin{bmatrix} \mathrm{sign}(\xi) \\ s \end{bmatrix}, \|s\|_\infty \leq 1 \right\}.$$

- $f(X) = \|X\|_\star$ and $X_0 = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^*$:

$$\partial f(\mathbf{x}_0) = \left\{ U \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} V^*, |d_i| \leq 1 \right\}.$$

- $f(x) = \|x\|_\infty$ and $x_0 = \begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix}$:

Subgradients

The subgradient of a *convex function* is a *convex set*. In most cases of interest subgradients are easy to compute. Here are some examples:

- $f(x) = \|x\|_1$ and $x_0 = \begin{bmatrix} \xi \\ 0 \end{bmatrix}$:

$$\partial f(\mathbf{x}_0) = \left\{ \begin{bmatrix} \text{sign}(\xi) \\ s \end{bmatrix}, \|s\|_\infty \leq 1 \right\}.$$

- $f(X) = \|X\|_\star$ and $X_0 = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^*$:

$$\partial f(\mathbf{x}_0) = \left\{ U \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} V^*, |d_i| \leq 1 \right\}.$$

- $f(x) = \|x\|_\infty$ and $x_0 = \begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix}$:

$$\partial f(\mathbf{x}_0) = \left\{ \begin{bmatrix} s \\ -t \end{bmatrix}, s \geq 0, t \geq 0, \|s\|_1 + \|t\|_1 \leq 1 \right\}.$$

Squared Error of Generalized LASSO $\sigma \to 0$

Returning back to the (AO):

$$\min_{w} \sqrt{\|w\|^2 + \sigma^2}\|g\| + h_w^T w + h_\sigma \sigma + \lambda \sup_{s \in \partial f(\mathbf{x}_0)} s^T(-w),$$

Squared Error of Generalized LASSO $\sigma \to 0$

Returning back to the (AO):

$$\min_{w} \sqrt{\|w\|^2 + \sigma^2}\|g\| + h_w^T w + h_\sigma \sigma + \lambda \sup_{s \in \partial f(\mathbf{x}_0)} s^T(-w),$$

or

$$\min_{w} \sqrt{\|w\|^2 + \sigma^2}\|g\| + \sup_{s \in \lambda \partial f(\mathbf{x}_0)} (h_w - s)^T w.$$

Squared Error of Generalized LASSO $\sigma \to 0$

Returning back to the (AO):

$$\min_{w} \sqrt{\|w\|^2 + \sigma^2}\|g\| + h_w^T w + h_\sigma \sigma + \lambda \sup_{s \in \partial f(\mathbf{x}_0)} s^T(-w),$$

or

$$\min_{w} \sqrt{\|w\|^2 + \sigma^2}\|g\| + \sup_{s \in \lambda\partial f(\mathbf{x}_0)} (h_w - s)^T w.$$

As before, fixing the norm $\|w\| = \alpha$, optimization over the direction of w is straightforward:

$$\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| + \sup_{s \in \lambda\partial f(\mathbf{x}_0)} -\alpha\|h_w - s\|.$$

Squared Error of Generalized LASSO $\sigma \to 0$

Returning back to the (AO):

$$\min_w \sqrt{\|w\|^2 + \sigma^2}\|g\| + h_w^T w + h_\sigma \sigma + \lambda \sup_{s \in \partial f(\mathbf{x}_0)} s^T(-w),$$

or

$$\min_w \sqrt{\|w\|^2 + \sigma^2}\|g\| + \sup_{s \in \lambda \partial f(\mathbf{x}_0)} (h_w - s)^T w.$$

As before, fixing the norm $\|w\| = \alpha$, optimization over the direction of w is straightforward:

$$\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| + \sup_{s \in \lambda \partial f(\mathbf{x}_0)} -\alpha \|h_w - s\|.$$

Or:

$$\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| - \alpha \underbrace{\inf_{s \in \lambda \partial f(\mathbf{x}_0)} \|h_w - s\|}_{\text{dist}(h_w, \lambda \partial f(\mathbf{x}_0))}.$$

Squared Error of Generalized LASSO $\sigma \to 0$

$$\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| - \alpha \mathrm{dist}(h_w, \lambda \partial f(\mathbf{x}_0)).$$

Squared Error of Generalized LASSO $\sigma \to 0$

$$\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| - \alpha \mathrm{dist}(h_w, \lambda \partial f(\mathbf{x}_0)).$$

This looks exactly like what we had for least-squares: $\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| - \alpha \|h_w\|$.

Squared Error of Generalized LASSO $\sigma \to 0$

$$\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| - \alpha \text{dist}(h_w, \lambda \partial f(\mathbf{x}_0)).$$

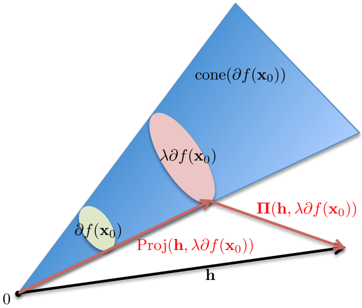
This looks exactly like what we had for least-squares: $\min_{\alpha \geq 0} \sqrt{\alpha^2 + \sigma^2}\|g\| - \alpha\|h_w\|$. Differentiating over α yields:

$$\lim_{\sigma \to 0} \frac{\alpha^2}{\sigma^2} = \frac{\text{dist}^2(h_w, \lambda \partial f(\mathbf{x}_0))}{m - \text{dist}^2(h_w, \lambda \partial f(\mathbf{x}_0))}.$$

Main Result: The Squared Error of Generalized LASSO

Generate an n-dimensional vector h with iid $N(0,1)$ entries and define:

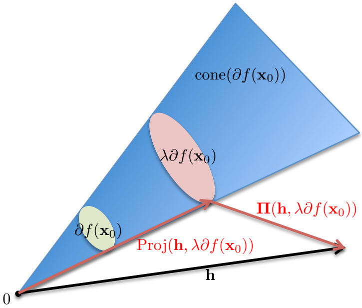
$$D_f(x_0,\lambda) = E\text{dist}^2\left(h, \lambda\partial f(x_0)\right).$$



Main Result: The Squared Error of Generalized LASSO

Generate an n-dimensional vector h with iid $N(0,1)$ entries and define:

$$D_f(x_0,\lambda) = E\text{dist}^2\left(h, \lambda\partial f(x_0)\right).$$



It turns out that $\text{dist}^2(h_w, \lambda\partial f(\mathbf{x}_0))$ concentrates to $D_f(x_0,\lambda)$, and that:

$$\lim_{\sigma\to 0} \frac{\|x_0 - \hat{x}\|^2}{\sigma^2} \to \frac{D_f(x_0,\lambda)}{m - D_f(x_0,\lambda)}.$$

Main Result

$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\sigma^2} \to \frac{D_f(x_0, \lambda)}{m - D_f(x_0, \lambda)}.$$

Main Result

$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\sigma^2} \to \frac{D_f(x_0, \lambda)}{m - D_f(x_0, \lambda)}.$$

- Note that, compared to the normalized mean-square error of standard least-squares, $\frac{n}{m-n}$, the ambient dimension n has been replaced by $D_f(x_0, \lambda)$.

Main Result

$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\sigma^2} \to \frac{D_f(x_0, \lambda)}{m - D_f(x_0, \lambda)}.$$

- Note that, compared to the normalized mean-square error of standard least-squares, $\frac{n}{m-n}$, the ambient dimension n has been replaced by $D_f(x_0, \lambda)$.
- The value of λ that minimizes the mean-square error is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda).$$

Main Result

$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\sigma^2} \to \frac{D_f(x_0, \lambda)}{m - D_f(x_0, \lambda)}.$$

- Note that, compared to the normalized mean-square error of standard least-squares, $\frac{n}{m-n}$, the ambient dimension n has been replaced by $D_f(x_0, \lambda)$.
- The value of λ that minimizes the mean-square error is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda).$$

It is easy to see that

$$D_f(x_0, \lambda^*) = E\text{dist}^2\left(h, \text{cone}(\partial f(x_0))\right) \stackrel{\Delta}{=} \omega^2.$$

Main Result

- $$\omega^2 = E\text{dist}^2\left(h, \text{cone}(\partial f(x_0))\right)$$

  The quantity ω^2 is the squared *Gaussian width* of the cone of the subgradient and has been referred to as the *statistical dimension* by Tropp et al.

Main Result

- $$\omega^2 = E\text{dist}^2\left(h, \text{cone}(\partial f(x_0))\right)$$

  The quantity ω^2 is the squared *Gaussian width* of the cone of the subgradient and has been referred to as the *statistical dimension* by Tropp et al.

- Thus, for the optimum choice of λ:

$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{\omega^2}{m - \omega^2}.$$

Main Result

- $$\omega^2 = E\text{dist}^2\left(h, \text{cone}(\partial f(x_0))\right)$$

  The quantity ω^2 is the squared *Gaussian width* of the cone of the subgradient and has been referred to as the *statistical dimension* by Tropp et al.

- Thus, for the optimum choice of λ:

$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{\omega^2}{m - \omega^2}.$$

- The quantity ω^2 determines the minimum number of measurements required to recover a k-sparse signal using (appropriate) convex optimization. (The so-called *recovery thresholds*.)

Statistical Dimension

- The quantity $D_f(x_0, \lambda)$ is easy to numerically compute and ω^2 can often be computed in closed form.

Statistical Dimension

- The quantity $D_f(x_0, \lambda)$ is easy to numerically compute and ω^2 can often be computed in closed form.
- For n-dimensional k-sparse signals and $f(x) = \|x\|_1$:

$$\omega^2 = 2k \log \frac{2n}{k} \quad , \quad \lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{2k \log \frac{2n}{k}}{m - 2k \log \frac{2n}{k}}$$

Statistical Dimension

- The quantity $D_f(x_0, \lambda)$ is easy to numerically compute and ω^2 can often be computed in closed form.
- For n-dimensional k-sparse signals and $f(x) = \|x\|_1$:

$$\omega^2 = 2k \log \frac{2n}{k} \quad , \quad \lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{2k \log \frac{2n}{k}}{m - 2k \log \frac{2n}{k}}$$

- For $n \times n$ rank r matrices and $f(X) = \|X\|_\star$:

$$\omega^2 = 3r(2n - r) \quad , \quad \lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{3r(2n - r)}{m - 3r(2n - r)}$$

Statistical Dimension

- The quantity $D_f(x_0, \lambda)$ is easy to numerically compute and ω^2 can often be computed in closed form.
- For n-dimensional k-sparse signals and $f(x) = \|x\|_1$:

$$\omega^2 = 2k \log \frac{2n}{k} \quad , \quad \lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{2k \log \frac{2n}{k}}{m - 2k \log \frac{2n}{k}}$$

- For $n \times n$ rank r matrices and $f(X) = \|X\|_\star$:

$$\omega^2 = 3r(2n - r) \quad , \quad \lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{3r(2n - r)}{m - 3r(2n - r)}$$

- For BPSK signals and $f(x) = \|x\|_\infty$:

$$\omega^2 = \frac{n}{2} \quad , \quad \lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{n/2}{m - n/2} = \frac{n}{2m - n}$$

Example

$\mathbf{X}_0 \in \mathbb{R}^{n\times n}$ is rank r. Observe, $\mathbf{y} = A\cdot\text{vec}(X_0) + \mathbf{z}$, solve the Matrix LASSO,

$$\min_{\mathbf{X}} \{\|\mathbf{y} - A\cdot\text{vec}(X)\|_2 + \lambda\|\mathbf{X}\|_\star\}$$

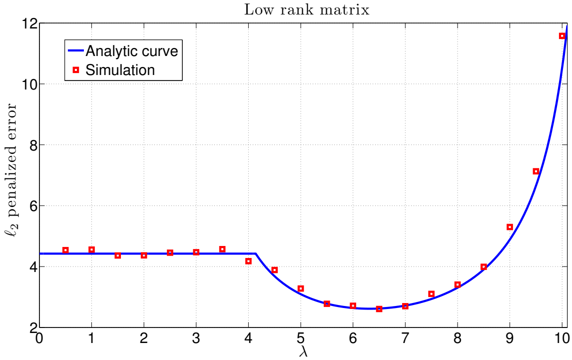


Figure: $n = 45$, $r = 6$, measurements $m = 0.6n^2$.

Phase Transitions for Convex Relaxation - Some History

- In the ℓ_1 case the subgradient cone is polyhedral and Donoho and Tanner (2005) computed the Grassman angle to obtain the minimum number of measurements required to recover a k-sparse signal
  - very cumbersome calculations, required considering exponentially many inner and outer angles, etc.

Phase Transitions for Convex Relaxation - Some History

- In the ℓ_1 case the subgradient cone is polyhedral and Donoho and Tanner (2005) computed the Grassman angle to obtain the minimum number of measurements required to recover a k-sparse signal
  - very cumbersome calculations, required considering exponentially many inner and outer angles, etc.
- Extended to robustness and weighted ℓ_1 by Xu-H in 2007 (even more cumbersome)

Phase Transitions for Convex Relaxation - Some History

- In the ℓ_1 case the subgradient cone is polyhedral and Donoho and Tanner (2005) computed the Grassman angle to obtain the minimum number of measurements required to recover a k-sparse signal
  - very cumbersome calculations, required considering exponentially many inner and outer angles, etc.
- Extended to robustness and weighted ℓ_1 by Xu-H in 2007 (even more cumbersome)
- Donoho-Tanner approach hard to extend (Recht-Xu-H (2008) attempted this for nuclear norm—only obtained bounds since subgradient cone is non-polyhedral)

Phase Transitions for Convex Relaxation - Some History

- In the ℓ_1 case the subgradient cone is polyhedral and Donoho and Tanner (2005) computed the Grassman angle to obtain the minimum number of measurements required to recover a k-sparse signal
  - very cumbersome calculations, required considering exponentially many inner and outer angles, etc.
- Extended to robustness and weighted ℓ_1 by Xu-H in 2007 (even more cumbersome)
- Donoho-Tanner approach hard to extend (Recht-Xu-H (2008) attempted this for nuclear norm—only obtained bounds since subgradient cone is non-polyhedral)
- New framework developed by Rudelson and Vershynin (2006) and, especially, Stojnic in 2009 (using escape-through-mesh and Gaussian widths)

Phase Transitions for Convex Relaxation - Some History

- In the ℓ_1 case the subgradient cone is polyhedral and Donoho and Tanner (2005) computed the Grassman angle to obtain the minimum number of measurements required to recover a k-sparse signal
  - very cumbersome calculations, required considering exponentially many inner and outer angles, etc.
- Extended to robustness and weighted ℓ_1 by Xu-H in 2007 (even more cumbersome)
- Donoho-Tanner approach hard to extend (Recht-Xu-H (2008) attempted this for nuclear norm—only obtained bounds since subgradient cone is non-polyhedral)
- New framework developed by Rudelson and Vershynin (2006) and, especially, Stojnic in 2009 (using escape-through-mesh and Gaussian widths)
  - rederived results for sparse vectors; new results for block-sparse vectors

Phase Transitions for Convex Relaxation - Some History

- In the ℓ_1 case the subgradient cone is polyhedral and Donoho and Tanner (2005) computed the Grassman angle to obtain the minimum number of measurements required to recover a k-sparse signal
  - very cumbersome calculations, required considering exponentially many inner and outer angles, etc.
- Extended to robustness and weighted ℓ_1 by Xu-H in 2007 (even more cumbersome)
- Donoho-Tanner approach hard to extend (Recht-Xu-H (2008) attempted this for nuclear norm—only obtained bounds since subgradient cone is non-polyhedral)
- New framework developed by Rudelson and Vershynin (2006) and, especially, Stojnic in 2009 (using escape-through-mesh and Gaussian widths)
  - rederived results for sparse vectors; new results for block-sparse vectors
  - much simpler derivation

Phase Transitions for Convex Relaxation - Some History

Stojnic's new approach:

- Allowed the development of a general framework (Chandrasekaran-Parrilo-Willsky, 2010)
    - exact calculation for nuclear norm (Oymak-H, 2010)

Phase Transitions for Convex Relaxation - Some History

Stojnic's new approach:

- Allowed the development of a general framework (Chandrasekaran-Parrilo-Willsky, 2010)
  - exact calculation for nuclear norm (Oymak-H, 2010)
- Deconvolution (McCoy-Tropp, 2012)

Phase Transitions for Convex Relaxation - Some History

Stojnic's new approach:

- Allowed the development of a general framework (Chandrasekaran-Parrilo-Willsky, 2010)
  - exact calculation for nuclear norm (Oymak-H, 2010)
- Deconvolution (McCoy-Tropp, 2012)
- Tightness of Gaussian widths Stojnic, 2013 (for ℓ_1), Amelunxen-Lotz-McCoy-Tropp, 2013 (for the general case)

Phase Transitions for Convex Relaxation - Some History

Stojnic's new approach:

- Allowed the development of a general framework (Chandrasekaran-Parrilo-Willsky, 2010)
  - exact calculation for nuclear norm (Oymak-H, 2010)
- Deconvolution (McCoy-Tropp, 2012)
- Tightness of Gaussian widths Stojnic, 2013 (for ℓ_1), Amelunxen-Lotz-McCoy-Tropp, 2013 (for the general case)

Replica-based analysis:

- Guo, Baron and Shamai (2009), Kabashima, Wadayama, Tanaka (2009), Rangan, Fletecher, Goyal (2012), Vehkapera, Kabashima, Chatterjee (2013), Wen, Zhang, Wong, Chen (2014)

What About the Noisy Case?

- Noisy case for l_1 LASSO first studied by Bayati, Montanari and Donoho (2012) using approximate message passing

What About the Noisy Case?

- Noisy case for l_1 LASSO first studied by Bayati, Montanari and Donoho (2012) using approximate message passing
- A new approach developed by Stojnic (2013)

What About the Noisy Case?

- Noisy case for l_1 LASSO first studied by Bayati, Montanari and Donoho (2012) using approximate message passing
- A new approach developed by Stojnic (2013)
- Our approach is inspired by Stojnic (2013)

What About the Noisy Case?

- Noisy case for l_1 LASSO first studied by Bayati, Montanari and Donoho (2012) using approximate message passing
- A new approach developed by Stojnic (2013)
- Our approach is inspired by Stojnic (2013)
    - subsumes all earlier (noiseless and noisy results)
    - allows for much, much more (as we have seen and shall further see)
    - is the most natural way to study the problem

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say.

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say. This is usually not available.

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say. This is usually not available.

Question: How to tune λ?

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say. This is usually not available.

Question: How to tune λ?

Answer: Here is one possibility that uses the fact that $\phi(g, h) \approx \sigma\sqrt{m - D_f(x_0, \lambda)}$:

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say. This is usually not available.

Question: How to tune λ?

Answer: Here is one possibility that uses the fact that $\phi(g, h) \approx \sigma\sqrt{m - D_f(x_0, \lambda)}$:

1. Choose a λ and solve the l_1 LASSO.

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say. This is usually not available.

Question: How to tune λ?

Answer: Here is one possibility that uses the fact that $\phi(g, h) \approx \sigma\sqrt{m - D_f(x_0, \lambda)}$:

1. Choose a λ and solve the l_1 LASSO.
2. Find the numerical value of the optimal cost, C, say.

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say. This is usually not available.

Question: How to tune λ?

Answer: Here is one possibility that uses the fact that $\phi(g, h) \approx \sigma\sqrt{m - D_f(x_0, \lambda)}$:

1. Choose a λ and solve the l_1 LASSO.
2. Find the numerical value of the optimal cost, C, say.
3. Find the sparsity k such that

$$|C - \sigma\sqrt{m - D_f(x_0, \lambda)}|,$$

   is minimized.

Tuning the Regularizer λ

The optimal value of λ is given by

$$\lambda^* = \arg\min_{\lambda \geq 0} D_f(x_0, \lambda),$$

which requires knowledge of the sparsity of x_0, say. This is usually not available.

Question: How to tune λ?

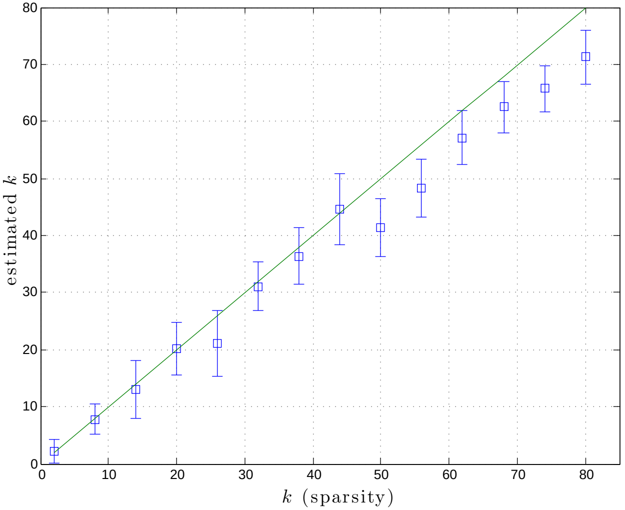
Answer: Here is one possibility that uses the fact that $\phi(g, h) \approx \sigma\sqrt{m - D_f(x_0, \lambda)}$:

1. Choose a λ and solve the l_1 LASSO.
2. Find the numerical value of the optimal cost, C, say.
3. Find the sparsity k such that

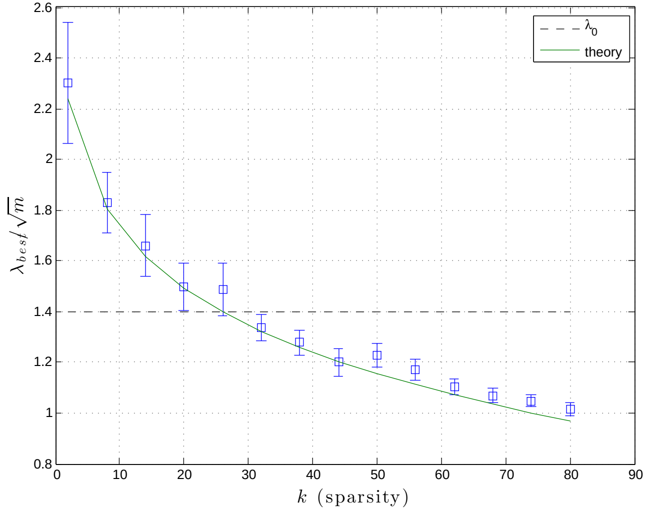
$$|C - \sigma\sqrt{m - D_f(x_0, \lambda)}|,$$

   is minimized.
4. For this value of k find the optimal λ^*.

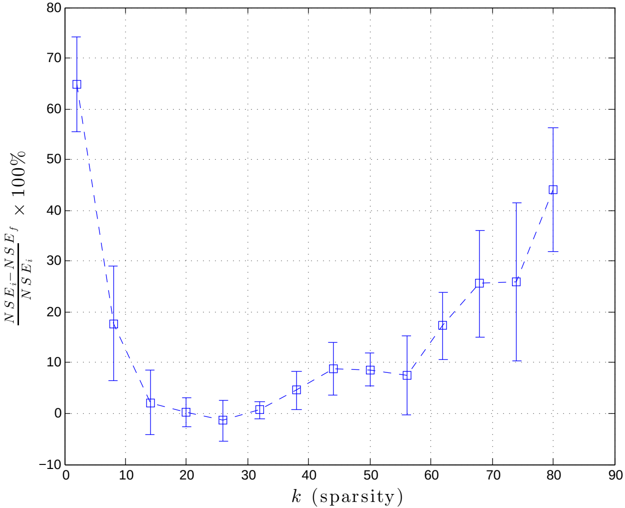
Estimating the Sparsity: $n = 520$, $m = 280$



Tuning λ: $n = 520$, $m = 280$



Improvement in NSE: $n = 520$, $m = 280$



Generalizations

Finite σ and General Loss Functions

In the general case, the problem to study is:

$$\hat{x} = \arg\min_x \mathcal{L}(y - Ax) + \lambda f(x).$$

Finite σ and General Loss Functions

In the general case, the problem to study is:

$$\hat{x} = \arg\min_x \mathcal{L}(y - Ax) + \lambda f(x).$$

To turn this into a PO it is useful to rewrite $\mathcal{L}(\cdot)$ and $f(\cdot)$ in terms of their *Fenchel duals*

$$\mathcal{L}(y - Ax) = \max_u u^T(y - Ax) - \mathcal{L}^*(u) \quad \text{and} \quad f(x) = \max_v v^T x - f^*(v),$$

to obtain

$$\min_x \max_{u,v} u^T(y - Ax) - \mathcal{L}^*(u) + \lambda v^T x - \lambda f^*(v).$$

Finite σ and General Loss Functions

In the general case, the problem to study is:

$$\hat{x} = \arg\min_x \mathcal{L}(y - Ax) + \lambda f(x).$$

To turn this into a PO it is useful to rewrite $\mathcal{L}(\cdot)$ and $f(\cdot)$ in terms of their *Fenchel duals*

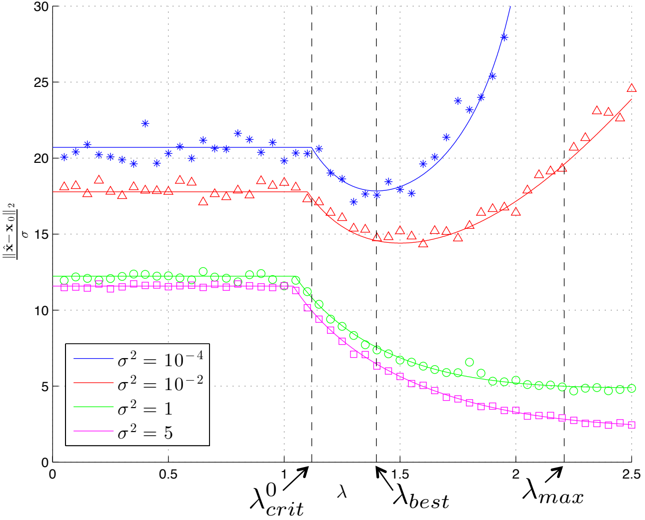
$$\mathcal{L}(y - Ax) = \max_u u^T(y - Ax) - \mathcal{L}^*(u) \quad \text{and} \quad f(x) = \max_v v^T x - f^*(v),$$

to obtain

$$\min_x \max_{u,v} u^T(y - Ax) - \mathcal{L}^*(u) + \lambda v^T x - \lambda f^*(v).$$

It turns out that the geometric quantities that show up in the analysis of the AO are the *expected Moreau envelopes*.

NSE for Finite σ: $n = 500$, $m = 150$, $k = 20$



Another Example: Least-Absolute Deviations (LAD)

We can do other loss functions.

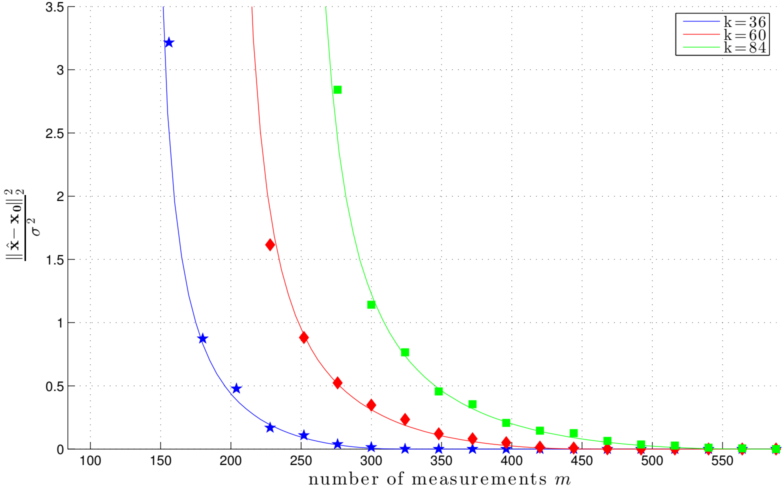
Another Example: Least-Absolute Deviations (LAD)

We can do other loss functions. For example,

$$\hat{x} = \arg\min_{x} \|y - Ax\|_1 + \lambda\|x\|_1,$$

which attempts to find a sparse signal in sparse noise and which is called *least absolute deviations* (LAD).

Squared Error vs Number of Measurements



Universality

- Our results assumed an iid Gaussian A.

Universality

- Our results assumed an iid Gaussian A.
- Is this necessary?

Universality

- Our results assumed an iid Gaussian A.
- Is this necessary?
- Simulations suggest that any iid distribution with the same second order statistics works.

Universality

- Our results assumed an iid Gaussian A.
- Is this necessary?
- Simulations suggest that any iid distribution with the same second order statistics works.
- We have been able to prove this for quadratic loss functions (OTTH 2015).

Universality

- Our results assumed an iid Gaussian A.
- Is this necessary?
- Simulations suggest that any iid distribution with the same second order statistics works.
- We have been able to prove this for quadratic loss functions (OTTH 2015). The value

$$\min_x \|y - Ax\|_2 + \lambda f(x),$$

concentrates for *any* A with iid zero-mean unit variance entries.

Universality

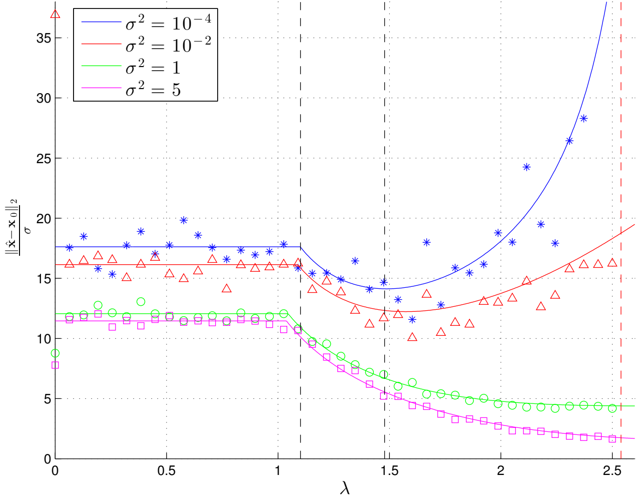
- Our results assumed an iid Gaussian A.
- Is this necessary?
- Simulations suggest that any iid distribution with the same second order statistics works.
- We have been able to prove this for quadratic loss functions (OTTH 2015). The value

$$\min_x \|y - Ax\|_2 + \lambda f(x),$$

concentrates for *any* A with iid zero-mean unit variance entries.

- Have yet to prove this for other loss functions and for the general (PO)

NSE for iid Bernouli($\frac{1}{2}$): $n = 500$, $m = 150$, $k = 20$



Other Matrix Ensembles - Haar

- Can we give results for non iid random matrix ensembles?

Other Matrix Ensembles - Haar

- Can we give results for non iid random matrix ensembles?
- An important class of random matrices are *isotropically random unitary matrices*,

Other Matrix Ensembles - Haar

- Can we give results for non iid random matrix ensembles?
- An important class of random matrices are *isotropically random unitary matrices*, i.e., matrices $Q \in R^{m \times n}$ $(m < n)$, such that

$$QQ^T = I_m, \qquad P(\Theta Q \Omega) = P(Q),$$

for all orthogonall Θ and Ω.

Other Matrix Ensembles - Haar

- Can we give results for non iid random matrix ensembles?
- An important class of random matrices are *isotropically random unitary matrices*, i.e., matrices $Q \in R^{m\times n}$ $(m < n)$, such that

$$QQ^T = I_m, \qquad P(\Theta Q \Omega) = P(Q),$$

  for all orthogonall Θ and Ω.
- For such random matrices, we have shown that the two optimization problems:

$$\begin{aligned} \Phi(Q,z) &= \min_w \quad \|\sigma z - Qw\| + \lambda f(w) && \text{(PO)} \\ \phi(g,h) &= \min_{w,l}\max_{\beta\geq 0} \quad \|\sigma v - w - l\| + \beta(\|l\|\cdot\|g\| - h^T l) + \lambda f(w) && \text{(AO)} \end{aligned}$$

  where z, v, h and g have iid $N(0,1)$ entries, have the same optimal costs and statistically the same optimal minimizer.

Isotropically Random Unitary Matrices

- Using the above result, we have been able to show that

$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{D_f(x_0, \lambda)}{m - D_f(x_0, \lambda)} \cdot \frac{n - D_f(x_0, \lambda)}{n}.$$

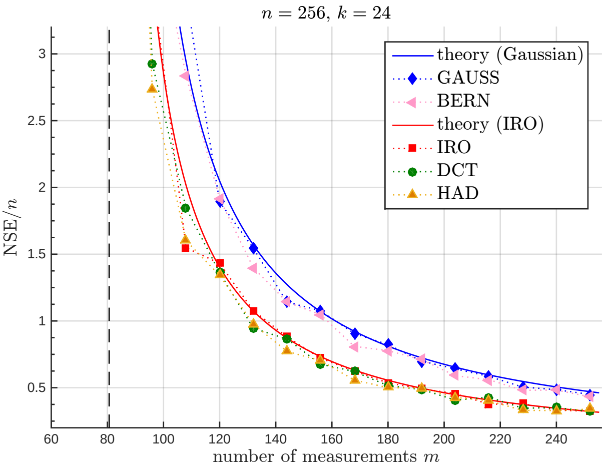
Isotropically Random Unitary Matrices

- Using the above result, we have been able to show that

$$\lim_{\sigma\to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{D_f(x_0,\lambda)}{m - D_f(x_0,\lambda)} \cdot \frac{n - D_f(x_0,\lambda)}{n}.$$

- Since $\frac{n - D_f(x_0,\lambda)}{n} < 1$, this is strictly better than the Gaussian case.

NSE for Isotropically Unitary Matrix: $n = 520$, $k = 20$



Nonlinear Measurements

Suppose we make nonlinear observations of the form

$$y = g(Ax_0 + v),$$

for some nonlinear function $g(\cdot)$.

Nonlinear Measurements

Suppose we make nonlinear observations of the form

$$y = g(Ax_0 + v),$$

for some nonlinear function $g(\cdot)$. For example, one-bit quantization corresponds to:

$$y = \text{sign}(Ax_0 + v).$$

Nonlinear Measurements

Suppose we make nonlinear observations of the form

$$y = g(Ax_0 + v),$$

for some nonlinear function $g(\cdot)$. For example, one-bit quantization corresponds to:

$$y = \mathrm{sign}(Ax_0 + v).$$

What happens if we apply generalized LASSO to such nonlinear measurements:

$$\min_x \|y - Ax\|_2 + \lambda f(x)?$$

Nonlinear Measurements

Suppose we make nonlinear observations of the form

$$y = g(Ax_0 + v),$$

for some nonlinear function $g(\cdot)$. For example, one-bit quantization corresponds to:

$$y = \mathrm{sign}(Ax_0 + v).$$

What happens if we apply generalized LASSO to such nonlinear measurements:

$$\min_x \|y - Ax\|_2 + \lambda f(x)?$$

This seems like a very naive thing to do.

Nonlinear Measurements

Suppose we make nonlinear observations of the form

$$y = g(Ax_0 + v),$$

for some nonlinear function $g(\cdot)$. For example, one-bit quantization corresponds to:

$$y = \text{sign}(Ax_0 + v).$$

What happens if we apply generalized LASSO to such nonlinear measurements:

$$\min_x \|y - Ax\|_2 + \lambda f(x)?$$

This seems like a very naive thing to do. However, it was suggested by Brillinger for standard least-squares in the 1980's and very recently by Plan and Vershynin.

Nonlinear Measurements

Theorem (TAH 2015): *The MSE of generalized LASSO for nonlinear measurements of the form* $y = g(Ax_0 + v)$ *is asymptotically the same as the MSE of generalized LASSO for measurements of the form* $y = \mu A x_0 + \sigma v$*, where:*

$$\mu = E\gamma g(\gamma) \quad \textit{and} \quad \sigma^2 = Eg^2(\gamma) - \mu^2 \quad \textit{for } \gamma \sim N(0,1).$$

Nonlinear Measurements

Theorem (TAH 2015): *The MSE of generalized LASSO for nonlinear measurements of the form $y = g(Ax_0 + v)$ is asymptotically the same as the MSE of generalized LASSO for measurements of the form $y = \mu A x_0 + \sigma v$, where:*

$$\mu = E\gamma g(\gamma) \quad \textit{and} \quad \sigma^2 = Eg^2(\gamma) - \mu^2 \quad \textit{for } \gamma \sim N(0,1).$$

- Therefore all the analysis we have done for generalized LASSO with linear measurements applies also to the nonlinear case.

Nonlinear Measurements

Theorem (TAH 2015): *The MSE of generalized LASSO for nonlinear measurements of the form $y = g(Ax_0 + v)$ is asymptotically the same as the MSE of generalized LASSO for measurements of the form $y = \mu A x_0 + \sigma v$, where:*

$$\mu = E\gamma g(\gamma) \quad \text{and} \quad \sigma^2 = Eg^2(\gamma) - \mu^2 \quad \text{for } \gamma \sim N(0,1).$$

- Therefore all the analysis we have done for generalized LASSO with linear measurements applies also to the nonlinear case.
- For 1-bit quantization we have:

$$\mu = \sqrt{\frac{2}{\pi}} \quad \text{and} \quad \sigma^2 = 1 - \frac{2}{\pi}$$

Nonlinear Measurements

Theorem (TAH 2015): *The MSE of generalized LASSO for nonlinear measurements of the form $y = g(Ax_0 + v)$ is asymptotically the same as the MSE of generalized LASSO for measurements of the form $y = \mu A x_0 + \sigma v$, where:*

$$\mu = E\gamma g(\gamma) \quad \text{and} \quad \sigma^2 = Eg^2(\gamma) - \mu^2 \quad \text{for } \gamma \sim N(0,1).$$

- Therefore all the analysis we have done for generalized LASSO with linear measurements applies also to the nonlinear case.
- For 1-bit quantization we have:

$$\mu = \sqrt{\frac{2}{\pi}} \quad \text{and} \quad \sigma^2 = 1 - \frac{2}{\pi}$$

- We can show that, for q-bit quantization, the optimal quantizer is the celebrated LLoyd-Max quantizer.

One-Bit Quantization

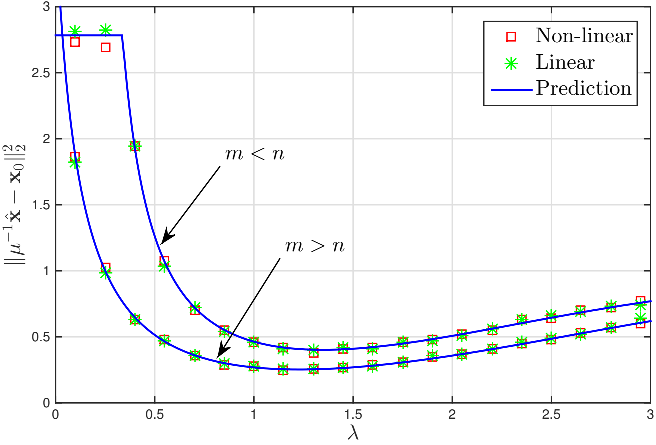


Figure: $n = 768$, $k = 115$, $m = 920 > n$ and $m = 576 < n$. The measurements were $y = \text{sign}(Ax_0 + .3v)$ with the v_i iid $N(0, 1)$.

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
  - *study an (AO) rather than the (PO)*

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
  - *study an (AO) rather than the (PO)*
- Allows for optimal tuning of regularizer parameters

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
  - *study an (AO) rather than the (PO)*
- Allows for optimal tuning of regularizer parameters
- Can consider various loss functions and regularizers

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
  - *study an (AO) rather than the (PO)*
- Allows for optimal tuning of regularizer parameters
- Can consider various loss functions and regularizers
- Results appear to be universal (proven for quadratic losses and general regularizers)

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
  - *study an (AO) rather than the (PO)*
- Allows for optimal tuning of regularizer parameters
- Can consider various loss functions and regularizers
- Results appear to be universal (proven for quadratic losses and general regularizers)
- Theory generalized to isotropically random unitary matrices

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
  - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
  - *study an (AO) rather than the (PO)*
- Allows for optimal tuning of regularizer parameters
- Can consider various loss functions and regularizers
- Results appear to be universal (proven for quadratic losses and general regularizers)
- Theory generalized to isotropically random unitary matrices
- Extends to nonlinear measurements

Summary and Conclusion

- Developed a general theory for the analysis of convex-based structured signal recovery problems for iid Gaussian measurement matrices
    - subsumes all known results (phase transtions, thresholds, etc.) and generates many new ones
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
    - *study an (AO) rather than the (PO)*
- Allows for optimal tuning of regularizer parameters
- Can consider various loss functions and regularizers
- Results appear to be universal (proven for quadratic losses and general regularizers)
- Theory generalized to isotropically random unitary matrices
- Extends to nonlinear measurements
- Generalization to quadratic Gaussian measurements would be very useful (for phase retrieval, graphical LASSO, etc.)