Recovering Structured Signals in Noise: Comparison Lemmas and the Performance of Convex Relaxation Methods

Babak Hassibi

California Institute of Technology

EUSIPCO, Nice, Cote d'Azur, France
August 31, 2015

Tribute to Dave Slepian (1923-2007)

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.
   DS: No, I am with a different group.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.
   DS: No, I am with a different group.
d. Will you be joining the French table?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.
   DS: No, I am with a different group.
d. Will you be joining the French table?
e. None of the above.

Outline

- **Introduction**
  - structured signal recovery
  - non-smooth convex optimization
  - LASSO and generalized LASSO
- **Comparison Lemmas**
  - Slepian, Gordon
- **Squared Error of Generalized LASSO**
  - Gaussian widths, statistical dimension
  - optimal parameter tuning
- **Generalizations**
  - other loss functions
  - other random matrix ensembles
- **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 such structured signal recovery methods

Non-Smooth Convex Optimization

- Non-smooth convex optimization has emerged as a tractable method to deal with such 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 such 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

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

Non-Smooth Convex Optimization

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

Non-Smooth Convex Optimization

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

- **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

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

- **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?)

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?)
  - 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?)
  - 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 *sparsest* 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 interesting.

Example: Noisy Compressed Sensing

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

Example: Noisy Compressed Sensing

Thus, the *information-theoretic* problem is perhaps not so 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 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

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

- 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*.

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*. But to do so, we need to tell an earlier story....

Example

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

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

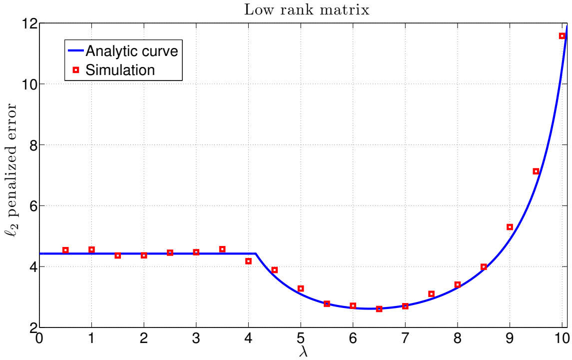


Figure: $n = 45$, $r = 6$, measurements $m = 0.6n^2$.

Noiseless 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 measurements of x using the $m \times n$ measurement matrix A to obtain

$$y = Ax.$$

Noiseless 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 measurements of x using the $m \times n$ measurement matrix A to obtain

$$y = Ax.$$

A heuristic (that has been around for decades) is:

$$\min \|x\|_1 \quad \text{subject to } y = Ax$$

Noiseless 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 measurements of x using the $m \times n$ measurement matrix A to obtain

$$y = Ax.$$

A heuristic (that has been around for decades) is:

$$\min \|x\|_1 \quad \text{subject to } y = Ax$$

The seminal work of Candes and Tao (2004) and Donoho (2004) has shown that *under certain conditions* the above ℓ_1 optimization can *exactly* recover the solution, thus avoiding an exponential search.

Noiseless 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 measurements of x using the $m \times n$ measurement matrix A to obtain

$$y = Ax.$$

A heuristic (that has been around for decades) is:

$$\min \|x\|_1 \quad \text{subject to } y = Ax$$

The seminal work of Candes and Tao (2004) and Donoho (2004) has shown that *under certain conditions* the above ℓ_1 optimization can *exactly* recover the solution, thus avoiding an exponential search.

- Candes and Tao showed that if A satisfies certain *restricted isometry* conditions, then ℓ_1 optimization works for small enough k
  - gives "order optimal", but **very** loose bounds

Exact Conditions for Signal Recovery

We will consider a general framework.

Consider a structured signal x_0, with a structure-inducing norm $f(\cdot) = \|\cdot\|$. We have access to *linear measurements* $y = \mathcal{A}(x_0) \in R^m$, and would like to know when we can recover the signal x_0 from the convex problem

$$\min \|x\| \quad \text{subject to } \mathcal{A}(x) = \mathcal{A}(x_0)?$$

Exact Conditions for Signal Recovery

We will consider a general framework.

Consider a structured signal x_0, with a structure-inducing norm $f(\cdot) = \|\cdot\|$. We have access to *linear measurements* $y = \mathcal{A}(x_0) \in R^m$, and would like to know when we can recover the signal x_0 from the convex problem

$$\min \|x\| \quad \text{subject to } \mathcal{A}(x) = \mathcal{A}(x_0)?$$

- For sparse signals we have the ℓ_1 norm; for nonuniform sparse signals the weighted ℓ_1 norm; for low rank matrices the nuclear norm

Exact Conditions for Signal Recovery

We will consider a general framework.

Consider a structured signal x_0, with a structure-inducing norm $f(\cdot) = \|\cdot\|$. We have access to *linear measurements* $y = \mathcal{A}(x_0) \in R^m$, and would like to know when we can recover the signal x_0 from the convex problem

$$\min \|x\| \quad \text{subject to } \mathcal{A}(x) = \mathcal{A}(x_0)?$$

- For sparse signals we have the ℓ_1 norm; for nonuniform sparse signals the weighted ℓ_1 norm; for low rank matrices the nuclear norm

Let $\mathcal{U}(x_0) = \{z, \|x_0 + z\| \leq \|x_0\|\}$. Then x_0 is the unique solution of the above convex problem iff:

Exact Conditions for Signal Recovery

We will consider a general framework.

Consider a structured signal x_0, with a structure-inducing norm $f(\cdot) = \|\cdot\|$. We have access to *linear measurements* $y = \mathcal{A}(x_0) \in R^m$, and would like to know when we can recover the signal x_0 from the convex problem

$$\min \|x\| \quad \text{subject to } \mathcal{A}(x) = \mathcal{A}(x_0)?$$

- For sparse signals we have the ℓ_1 norm; for nonuniform sparse signals the weighted ℓ_1 norm; for low rank matrices the nuclear norm

Let $\mathcal{U}(x_0) = \{z, \|x_0 + z\| \leq \|x_0\|\}$. Then x_0 is the unique solution of the above convex problem iff:

$$\mathcal{N}(\mathcal{A}) \cap \mathcal{U}(x_0) = \{0\}.$$

A Bit of Geometry: Subgradients and the Polar Cone

Note that $\mathcal{N}(\mathcal{A})$ is a linear subspace and that therefore the condition can be rewritten as

$$\mathcal{N}(\mathcal{A}) \cap \mathrm{cone}(\mathcal{U}(x_0)) = \{0\}.$$

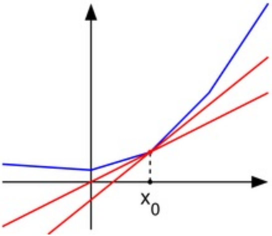
A Bit of Geometry: Subgradients and the Polar Cone

Note that $\mathcal{N}(\mathcal{A})$ is a linear subspace and that therefore the condition can be rewritten as

$$\mathcal{N}(\mathcal{A}) \cap \text{cone}(\mathcal{U}(x_0)) = \{0\}.$$

We can characterize $\text{cone}(\mathcal{U}(x_0))$ through the subgradient of the convex function $\|\cdot\|$:

$$\partial\|x_0\| = \{v | v^T(x - x_0) + \|x_0\| \leq \|x\|, \forall x\}.$$



A Bit of Geometry: Subgradients and the Polar Cone

It is now straightforward to see that

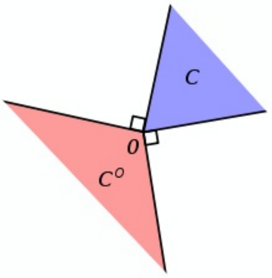
$$\text{cone}(\mathcal{U}(x_0)) = \{z | v^T z \le 0, \forall v \in \partial \|x_0\|\}.$$

A Bit of Geometry: Subgradients and the Polar Cone

It is now straightforward to see that

$$\text{cone}(\mathcal{U}(x_0)) = \{z | v^T z \leq 0, \forall v \in \partial\|x_0\|\}.$$

But this is simply the *polar cone* of $\partial\|x_0\|$.



Thus, we can recover x_0 from the convex problem iff:

$$\mathcal{N}(\mathcal{A}) \cap (\partial\|x_0\|)^O = \{0\}.$$

Phase Transitions for Exact Signal Recovery

- Thus, recovery depends on the null space of the measurement matrix and the polar cone of the subgradient (at the point we want to recover).

Phase Transitions for Exact Signal Recovery

- Thus, recovery depends on the null space of the measurement matrix and the polar cone of the subgradient (at the point we want to recover).
- While computing the polar cone of the subgradient is often straightforward, checking the condition $\mathcal{N}(\mathcal{A}) \cap (\partial\|x_0\|)^{O} = \{0\}$ for a *specific* $\mathcal{A}$ is difficult.

Phase Transitions for Exact Signal Recovery

- Thus, recovery depends on the null space of the measurement matrix and the polar cone of the subgradient (at the point we want to recover).
- While computing the polar cone of the subgradient is often straightforward, checking the condition $\mathcal{N}(\mathcal{A}) \cap (\partial\|x_0\|)^O = \{0\}$ for a *specific* $\mathcal{A}$ is difficult.
- Therefore the focus has been on checking whether the condition holds for a *family* of random $\mathcal{A}$'s with high probability.

Phase Transitions for Exact Signal Recovery

- Thus, recovery depends on the null space of the measurement matrix and the polar cone of the subgradient (at the point we want to recover).
- While computing the polar cone of the subgradient is often straightforward, checking the condition $\mathcal{N}(\mathcal{A}) \cap (\partial \|x_0\|)^O = \{0\}$ for a *specific* $\mathcal{A}$ is difficult.
- Therefore the focus has been on checking whether the condition holds for a *family* of random $\mathcal{A}$'s with high probability.
- It is customary to assume that the measurement matrix $\mathcal{A}$ is composed of iid zero-mean unit-variance entries.

Phase Transitions for Exact Signal Recovery

- Thus, recovery depends on the null space of the measurement matrix and the polar cone of the subgradient (at the point we want to recover).
- While computing the polar cone of the subgradient is often straightforward, checking the condition $\mathcal{N}(\mathcal{A}) \cap (\partial\|x_0\|)^O = \{0\}$ for a *specific* $\mathcal{A}$ is difficult.
- Therefore the focus has been on checking whether the condition holds for a *family* of random $\mathcal{A}$'s with high probability.
- It is customary to assume that the measurement matrix $\mathcal{A}$ is composed of iid zero-mean unit-variance entries.
- This makes the nullspace $\mathcal{N}(\mathcal{A})$ *rotationally-invariant*.

Phase Transitions for Exact Signal Recovery

- Thus, recovery depends on the null space of the measurement matrix and the polar cone of the subgradient (at the point we want to recover).
- While computing the polar cone of the subgradient is often straightforward, checking the condition $\mathcal{N}(\mathcal{A}) \cap (\partial\|x_0\|)^O = \{0\}$ for a *specific* $\mathcal{A}$ is difficult.
- Therefore the focus has been on checking whether the condition holds for a *family* of random $\mathcal{A}$'s with high probability.
- It is customary to assume that the measurement matrix $\mathcal{A}$ is composed of iid zero-mean unit-variance entries.
- This makes the nullspace $\mathcal{N}(\mathcal{A})$ *rotationally-invariant*.
- The probability that a rotationally-invariant subspace intersects a cone is called the *Grassman angle* of the cone.

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
  - is the most natural way to study the problem

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
    - is the most natural way to study the problem

Where does all this come from?

Tribute to Dave Slepian (1923-2007)

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.
   DS: No, I am with a different group.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.
   DS: No, I am with a different group.
d. Will you be joining the French table?

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.
   DS: No, I am with a different group.
d. Will you be joining the French table?
e. ✓None of the above.

Tribute to Dave Slepian (1923-2007)

David Slepian goes to a bar. What does the waitress say?

a. Claude just came in.
b. Will you be waiting for Jack?
c. Will you be attending the function?
   DS: What function?
   The prolate spheroidal wave function.
   DS: No, I am with a different group.
d. Will you be joining the French table?
e. ✓None of the above. *Would you care to compare our beers?*

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:

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

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:

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

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. $E(X_{ij} - \mu_{ij})(X_{ij'} - \mu_{ij'}) \le E(Y_{ij} - \mu_{ij})(Y_{ij'} - \mu_{ij'})$
2. $E(X_{ij} - \mu_{ij})(X_{i'j'} - \mu_{i'j'}) \ge E(Y_{ij} - \mu_{ij})(Y_{i'j'} - \mu_{i'j'})$

Then

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

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:

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

and

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

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:

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

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

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:

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

- 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 2014)

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

A Stronger Version of Gordon's Lemma (TOH 2014)

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

Theorem

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

A Stronger Version of Gordon's Lemma (TOH 2014)

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

Theorem

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

A Stronger Version of Gordon's Lemma (TOH 2014)

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

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*

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

A Stronger Version of Gordon's Lemma (TOH 2014)

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

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 optimizations over x are* strongly *convex, and $\phi(g,h)$ concentrates, then for* any *norm $\|\cdot\|$, for which $\|\hat{x}_\phi\|$ concentrates, with high probability we have*

$$\|\hat{x}_\Phi\| = \|\hat{x}_\phi\|\left(1 + o(1)\right)$$

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

$$m \geq n.$$

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

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

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,

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

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,

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

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

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

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

Least-Squares

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

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

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

The maximization over u is straightforward:

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

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

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

The maximization over u is straightforward:

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

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:

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

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 simpler optimization problem:

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

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 simpler optimization problem:

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

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

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

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

Squared Error of Generalized LASSO

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

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

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

Squared Error of Generalized LASSO

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

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

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

so that we obtain

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

Squared Error of Generalized LASSO

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

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

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

so that we obtain

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

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

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

Squared Error of Generalized LASSO

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

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

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

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

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

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:

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

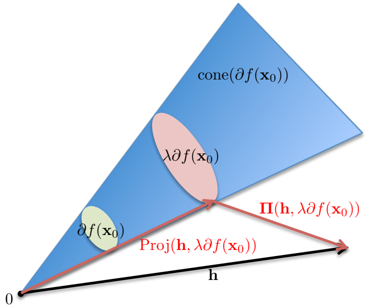
Differentiating over α yields:

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

Main Result: The Squared Error of Generalized LASSO

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

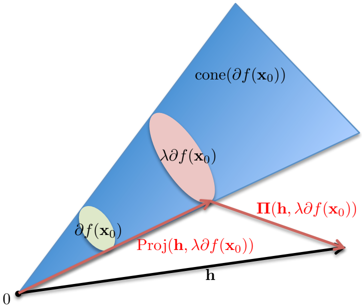
$$D_f(x_0, \lambda) = E\text{dist}^2\left(h, \lambda\partial f(x_0)\right).$$



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 $D_f(x_0,\lambda)$, so 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 qb-dimensional k block-sparse signals and $f(x) = \|x\|_{1,2}$:

$$\omega^2 = 4k(b + \log \frac{q}{k}) \quad , \quad \lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{4k(b + \log \frac{q}{k})}{m - 4k(b + \log \frac{q}{k})}$$

Example

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

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

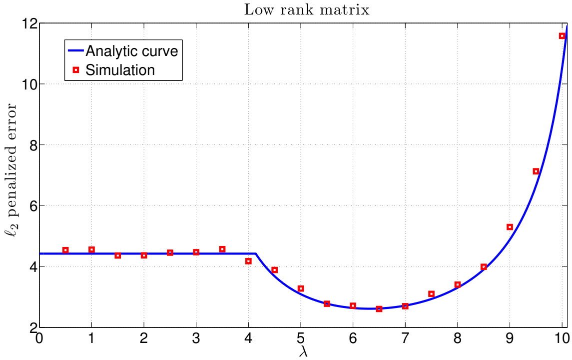


Figure: $n = 45$, $r = 6$, measurements $m = 0.6n^2$.

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?

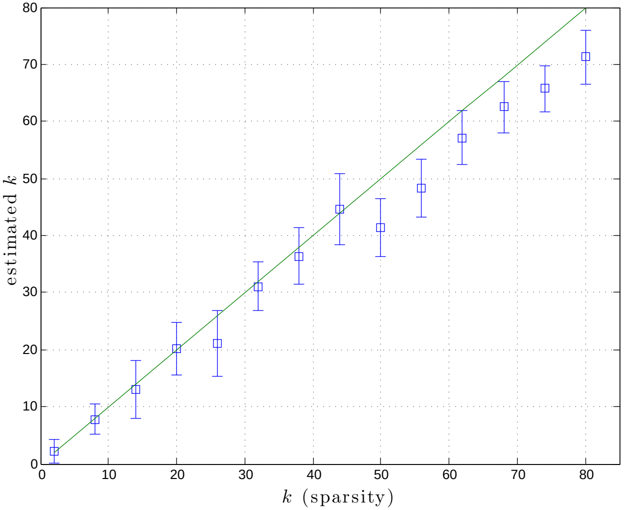
Answer: Here is one possibility that uses the fact that $\phi(g, h) \approx \sigma\sqrt{m - D_f(x_0, \lambda)}$:

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

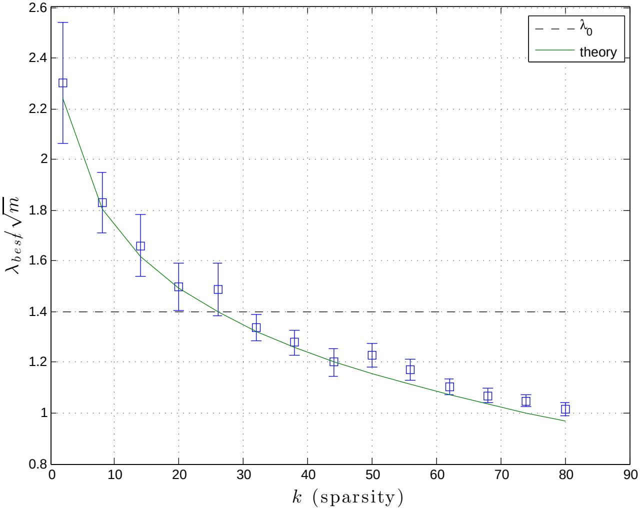
$$|C - \sigma\sqrt{m - D_f(x_0, \lambda)}|,$$

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

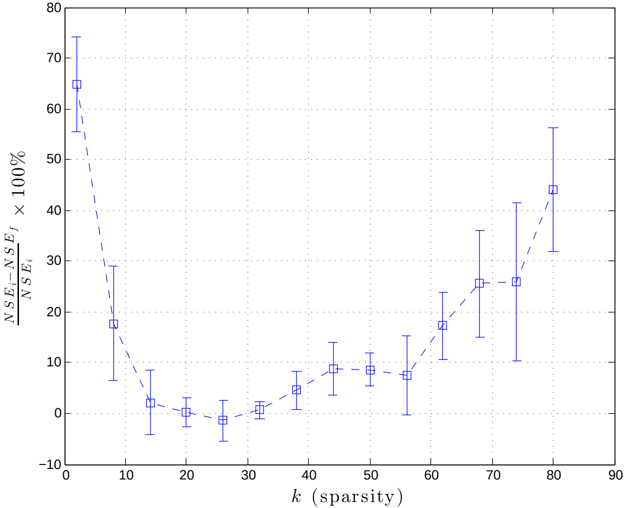
Estimating the Sparsity: $n = 520$, $m = 280$



Tuning λ: $n = 520$, $m = 280$



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



Generalizations

Finite σ

When σ is not very small, we must study:

$$\phi(g,h) = \min_{\mathbf{w}} \sqrt{\|\mathbf{w}\|^2 + \sigma^2}\|g\| - h^T w + \lambda\|x_0 + w\|_1.$$

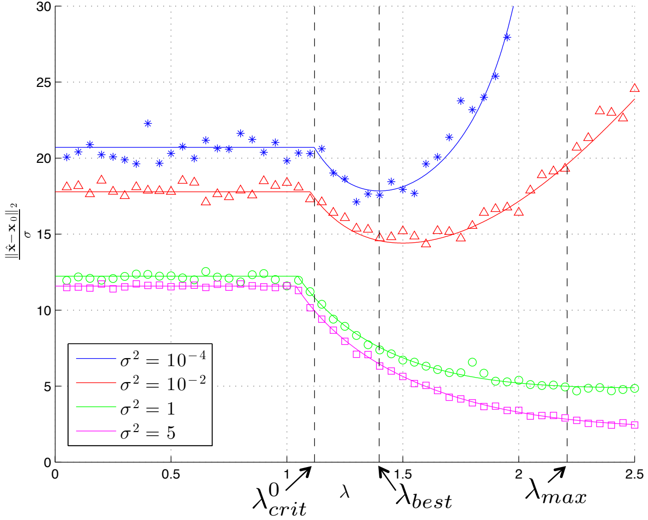
Finite σ

When σ is not very small, we must study:

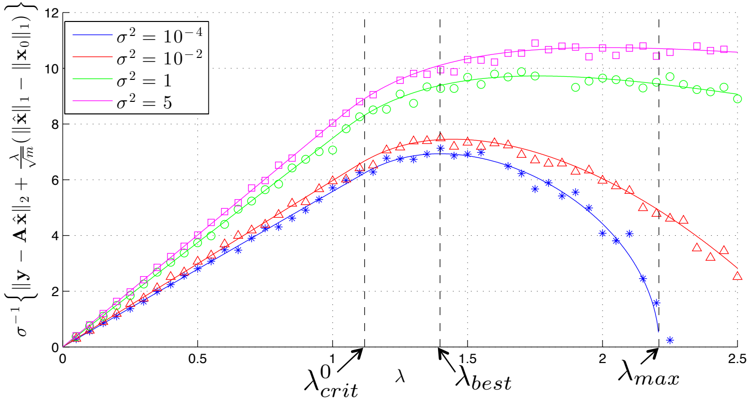
$$\phi(g,h) = \min_{\mathbf{w}} \sqrt{\|\mathbf{w}\|^2 + \sigma^2}\|g\| - h^T w + \lambda\|x_0 + w\|_1.$$

The analysis is a bit more complicated, but absolutely do-able.

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



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



Other Loss Functions

- We can do other loss functions.

Other Loss Functions

- We can do other loss functions. For example,

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

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

Other Loss Functions

- We can do other loss functions. For example,

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

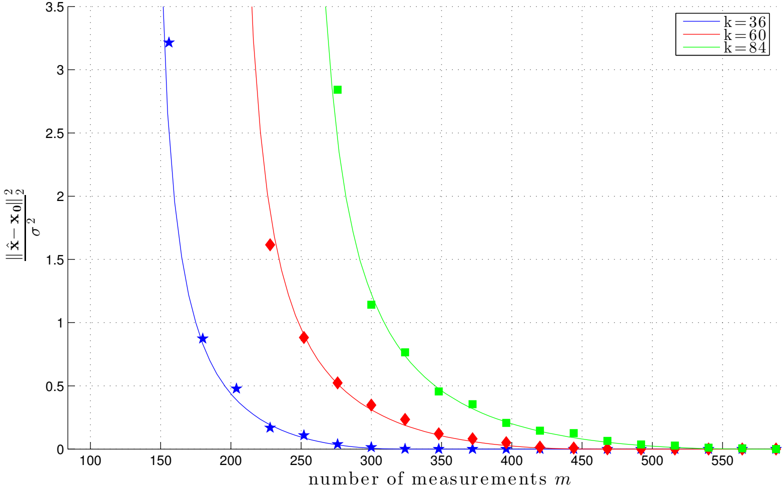
  which attempts to find a sparse signal in sparse noise and which is called *least absolute deviations* (LAD).

- In turns out that we now must analyze

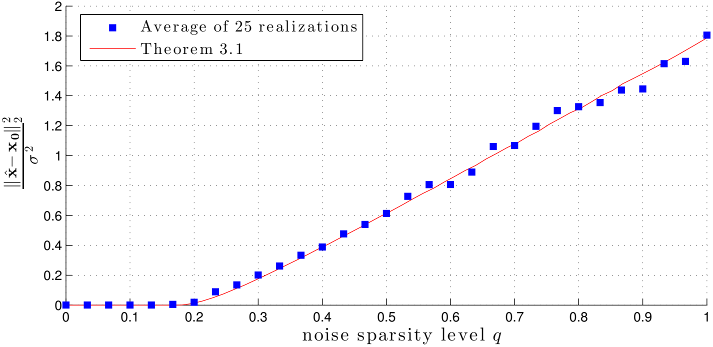
$$\phi(g,h) = \min_{\mathbf{w}} \max_{\|\mathbf{v}\|_\infty \le 1} \sqrt{\|\mathbf{w}\|^2 + \sigma^2} g^T \mathbf{v} - \|\mathbf{v}\| h^T \mathbf{w} + \sup_{\mathbf{s} \in \lambda \partial f(\mathbf{x}_0)} \mathbf{s}^T \mathbf{w}$$

  This is a bit more complicated, but still completely doable.

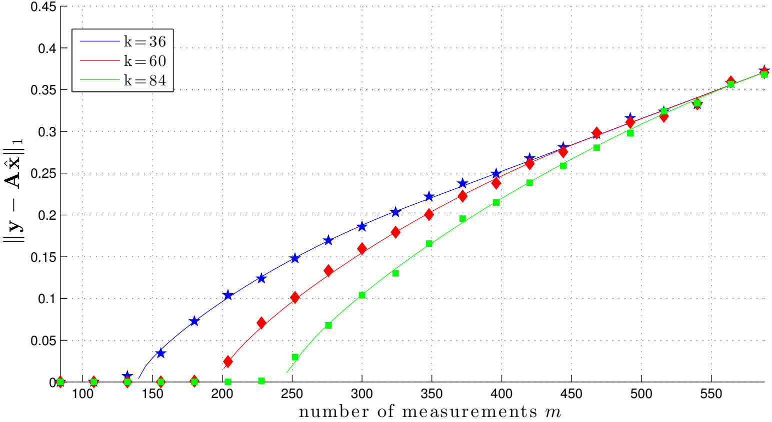
Squared Error vs Number of Measurements



Squared Error vs Sparsity of Noise



Cost vs Number of Measurements



Universality

- Our results assumed an iid Gaussian A.

Universality

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

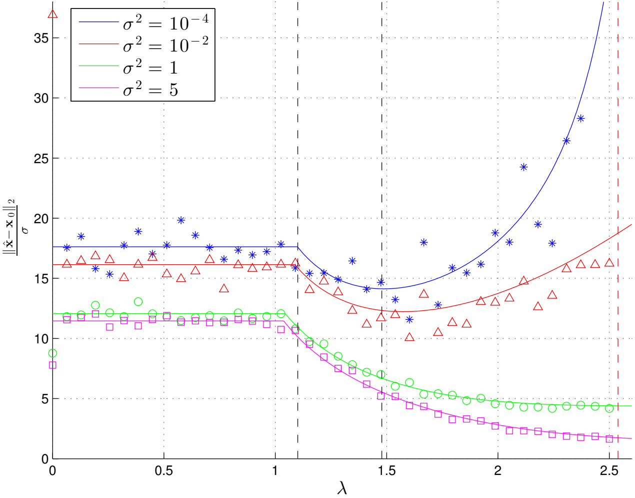
Universality

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

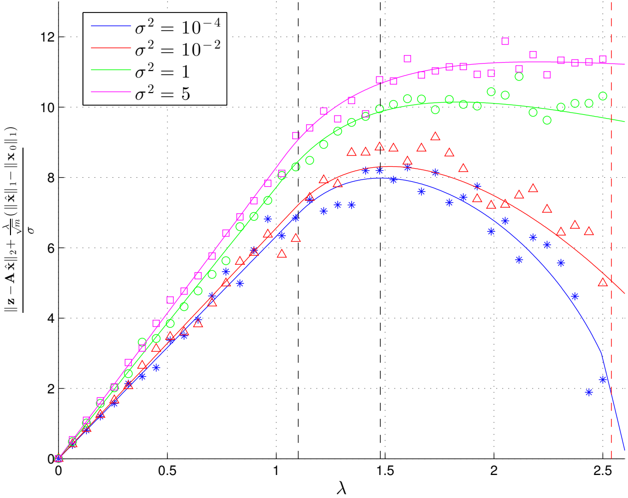
Universality

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

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



Cost for Bernoulli($\frac{1}{2}$): $n = 500$, $m = 150$, $k = 20$



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:

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

  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

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

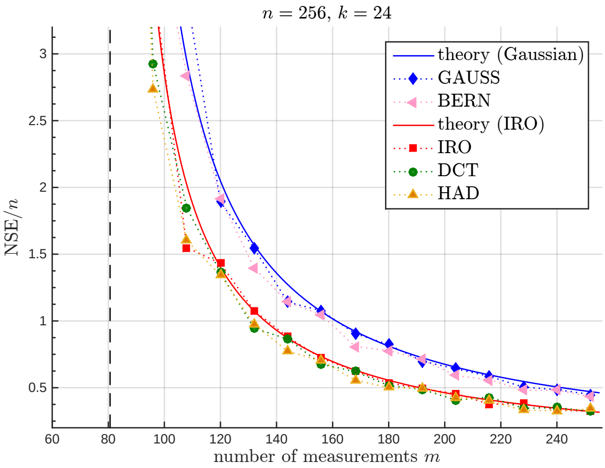
Isotropically Random Unitary Matrices

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

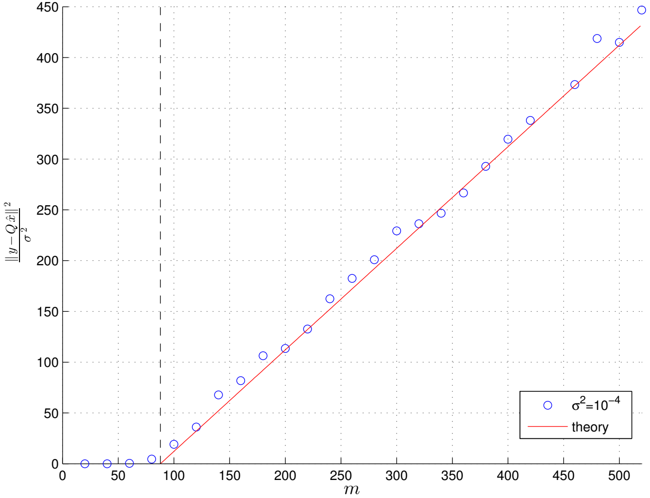
$$\lim_{\sigma \to 0} \frac{\|x_0 - \hat{x}\|^2}{\|z\|^2} \to \frac{D_f(x_0, \lambda)}{m - D_f(x_0, \lambda)} \cdot \frac{n - D_f(x_0, \lambda)}{n}.$$

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

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



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



Other Measurements: Quadratic Gaussians

In certain applications, such as *graphical LASSO* and *phase retrieval*, we encounter problems of the following form:

$$\min_{S \geq 0} \quad \mathrm{trace} G^T S G + \Psi(S), \tag{0.1}$$

where $S = S^T \in \mathcal{R}^{m \times m}$ and $G \in \mathcal{R}^{m \times n}$ has iid $N(0,1)$ entries.

Other Measurements: Quadratic Gaussians

In certain applications, such as *graphical LASSO* and *phase retrieval*, we encounter problems of the following form:

$$\min_{S \geq 0} \quad \text{trace} G^T S G + \Psi(S), \tag{0.1}$$

where $S = S^T \in \mathcal{R}^{m \times m}$ and $G \in \mathcal{R}^{m \times n}$ has iid $N(0,1)$ entries. For example, in graphical LASSO we have

$$\min_{S \geq 0} \quad \text{trace} G^T S G - N \log \det S + \lambda \|S\|_1.$$

Other Measurements: Quadratic Gaussians

In certain applications, such as *graphical LASSO* and *phase retrieval*, we encounter problems of the following form:

$$\min_{S \geq 0} \quad \mathrm{trace} G^T S G + \Psi(S), \tag{0.1}$$

where $S = S^T \in \mathcal{R}^{m \times m}$ and $G \in \mathcal{R}^{m \times n}$ has iid $N(0,1)$ entries. For example, in graphical LASSO we have

$$\min_{S \geq 0} \quad \mathrm{trace} G^T S G - N \log \det S + \lambda \|S\|_1.$$

Problem (0.1) can be *linearized* as:

$$\min_{S \geq 0} \max_{U} \quad 2\mathrm{trace} U^T S G - \mathrm{trace} U^T S U + \Psi(S).$$

Other Measurements: Quadratic Gaussians

In certain applications, such as *graphical LASSO* and *phase retrieval*, we encounter problems of the following form:

$$\min_{S \geq 0} \ \text{trace} G^T S G + \Psi(S), \tag{0.1}$$

where $S = S^T \in \mathcal{R}^{m \times m}$ and $G \in \mathcal{R}^{m \times n}$ has iid $N(0,1)$ entries. For example, in graphical LASSO we have

$$\min_{S \geq 0} \ \text{trace} G^T S G - N \log \det S + \lambda \|S\|_1.$$

Problem (0.1) can be *linearized* as:

$$\min_{S \geq 0} \max_{U} \ 2\text{trace} U^T S G - \text{trace} U^T S U + \Psi(S).$$

Can we come up with comparison lemmas for the Gaussian process $\text{trace} U^T S G$?

Summary and Conclusion

- Developed a general theory for the analysis of a wide range of structured signal recovery problems for iid Gaussian measurement matrices
- Theory builds on a strengthening of a lemma of Gordon (whose origin is one of Slepian)
- Allows for optimal tuning of regularizer parameter
- Various loss functions and regularizers can be considered
- Results appear to be universal ("close" to a proof)
- Theory generalized to isotropically random unitary matrices
- Generalization to quadratic Gaussian measurements would be very useful (for phase retrieval, graphical LASSO, etc.)