High-dimensional statistics: Some progress and challenges ahead

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

University College, London Master Class: Lecture 1

Introduction

- classical asymptotic theory: sample size $n \to +\infty$ with number of parameters $p$ fixed
  - law of large numbers, central limit theory
  - consistency of maximum likelihood estimation
Introduction

- classical asymptotic theory: sample size $n \to +\infty$ with number of parameters $p$ fixed
  - law of large numbers, central limit theory
  - consistency of maximum likelihood estimation

- modern applications in science and engineering:
  - large-scale problems: both $p$ and $n$ may be large (possibly $p \gg n$)
  - need for high-dimensional theory that provides non-asymptotic results for $(n, p)$
Introduction

- classical asymptotic theory: sample size $n \to +\infty$ with number of parameters $p$ fixed
  - law of large numbers, central limit theory
  - consistency of maximum likelihood estimation

- modern applications in science and engineering:
  - large-scale problems: both $p$ and $n$ may be large (possibly $p \gg n$)
  - need for high-dimensional theory that provides non-asymptotic results for $(n, p)$

- curses and blessings of high dimensionality
  - exponential explosions in computational complexity
  - statistical curses (sample complexity)
  - concentration of measure
Introduction

- modern applications in science and engineering:
  - large-scale problems: both \( p \) and \( n \) may be large (possibly \( p \gg n \))
  - need for high-dimensional theory that provides non-asymptotic results for \((n, p)\)

- curses and blessings of high dimensionality
  - exponential explosions in computational complexity
  - statistical curses (sample complexity)
  - concentration of measure

Key ideas:
- what embedded low-dimensional structures are present in data?
- how can they be exploited algorithmically?
Vignette I: High-dimensional matrix estimation

- want to estimate a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$
- given i.i.d. samples $X_i \sim N(0, \Sigma)$, for $i = 1, 2, \ldots, n$
want to estimate a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$

given i.i.d. samples $X_i \sim N(0, \Sigma)$, for $i = 1, 2, \ldots, n$

Classical approach:
Estimate $\Sigma$ via sample covariance matrix:

$$\hat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$$

average of $p \times p$ rank one matrices
Vignette I: High-dimensional matrix estimation

- want to estimate a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$
- given i.i.d. samples $X_i \sim N(0, \Sigma)$, for $i = 1, 2, \ldots, n$

Classical approach:
Estimate $\Sigma$ via sample covariance matrix:

$$\hat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$$

average of $p \times p$ rank one matrices

Reasonable properties: ($p$ fixed, $n$ increasing)

- Unbiased: $\mathbb{E}[\hat{\Sigma}_n] = \Sigma$
- Consistent: $\hat{\Sigma}_n \overset{a.s.}{\to} \Sigma$ as $n \to +\infty$
- Asymptotic distributional properties available
Vignette I: High-dimensional matrix estimation

- want to estimate a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$
- given i.i.d. samples $X_i \sim N(0, \Sigma)$, for $i = 1, 2, \ldots, n$

**Classical approach:**
Estimate $\Sigma$ via sample covariance matrix:

$$\hat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$$

average of $p \times p$ rank one matrices

**An alternative experiment:**
- Fix some $\alpha > 0$
- Study behavior over sequences with $\frac{p}{n} = \alpha$
- Does $\hat{\Sigma}_{n(p)}$ converge to anything reasonable?
Empirical vs MP law ($\alpha = 0.5$)

Empirical vs MP law ($\alpha = 0.2$)

Low-dimensional structure: Gaussian graphical models

Zero pattern of inverse covariance

\[ \mathbb{P}(x_1, x_2, \ldots, x_p) \propto \exp \left( - \frac{1}{2} x^T \Theta^* x \right). \]
Maximum-likelihood with $\ell_1$-regularization

Set-up: Samples from random vector with sparse covariance $\Sigma$ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{p \times p}$. 
Maximum-likelihood with $\ell_1$-regularization

Set-up: Samples from random vector with sparse covariance $\Sigma$ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{p \times p}$.

Estimator (for inverse covariance)

$$\hat{\Theta} \in \arg\min_{\Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T, \Theta \right\} - \log \det(\Theta) + \lambda_n \sum_{j \neq k} |\Theta_{jk}| \right\}$$

Some past work: Yuan & Lin, 2006; d’Asprémont et al., 2007; Bickel & Levina, 2007; El Karoui, 2007; d’Aspremont et al., 2007; Rothman et al., 2007; Zhou et al., 2007; Friedman et al., 2008; Lam & Fan, 2008; Ravikumar et al., 2008; Zhou, Cai & Huang, 2009
Problems with hidden variables: conditioned on hidden $Z$, vector $X = (X_1, X_2, X_3, X_4)$ is Gauss-Markov.
Gauss-Markov models with hidden variables

Problems with hidden variables: conditioned on hidden $Z$, vector $X = (X_1, X_2, X_3, X_4)$ is Gauss-Markov.

Inverse covariance of $X$ satisfies \{sparse, low-rank\} decomposition:

$$
\begin{bmatrix}
1 - \mu & \mu & \mu & \mu \\
\mu & 1 - \mu & \mu & \mu \\
\mu & \mu & 1 - \mu & \mu \\
\mu & \mu & \mu & 1 - \mu
\end{bmatrix} = I_{4 \times 4} - \mu 11^T.
$$

(Chandrasekaran, Parrilo & Willsky, 2010)
Vignette II: High-dimensional sparse linear regression

Set-up: noisy observations $y = X\theta^* + w$ with sparse $\theta^*$

Estimator: Lasso program

$$\hat{\theta} \in \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|$$

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Efron et al., 2004; Meinshausen & Buhlmann, 2005; Candes & Tao, 2005; Donoho, 2005; Haupt & Nowak, 2005; Zhou & Yu, 2006; Zou, 2006; Koltchinskii, 2007; van
Application A: Compressed sensing

(Donoho, 2005; Candes & Tao, 2005)

(a) Image: vectorize to $\beta^* \in \mathbb{R}^p$

(b) Compute $n$ random projections

$$y = X \times p$$
In practice, signals are sparse in a transform domain:

\[ \theta^* := \Psi \beta^* \] is a sparse signal,

where \( \Psi \) is an orthonormal matrix.

Reconstruct \( \theta^* \) (and hence image \( \beta^* = \Psi^T \theta^* \)) based on finding a sparse solution to under-constrained linear system

\[ y = \tilde{X} \theta \] where \( \tilde{X} = X \Psi^T \) is another random matrix.
Noiseless $\ell_1$ recovery: Unrescaled sample size

Prob. exact recovery vs. sample size ($\mu = 0$)

Probability of recovery versus sample size $n$. 
Application B: Graph structure estimation

- let $G = (V, E)$ be an undirected graph on $p = |V|$ vertices

- pairwise graphical model factorizes over edges of graph:
  \[
P(x_1, \ldots, x_p; \theta) \propto \exp \left\{ \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\}.
  \]

- given $n$ independent and identically distributed (i.i.d.) samples of $X = (X_1, \ldots, X_p)$, identify the underlying graph structure
Pseudolikelihood and neighborhood regression

- Markov properties encode neighborhood structure:

\[
\begin{align*}
(X_s \mid X_{V \backslash s}) & \quad \overset{d}{=} \quad (X_s \mid X_{N(s)}) \\
\text{Condition on full graph} & \quad \text{Condition on Markov blanket}
\end{align*}
\]

\[
N(s) = \{s, t, u, v, w\}
\]

- basis of pseudolikelihood method (Besag, 1974)
- basis of many graph learning algorithm (Friedman et al., 1999; Csiszar & Talata, 2005; Abeel et al., 2006; Meinshausen & Buhlmann, 2006)
Graph selection via neighborhood regression

<table>
<thead>
<tr>
<th>1 0 0 1 1 0 1</th>
<th>0 0 0 1 1 1 0 1</th>
<th>0 1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 1 0 0 0</td>
<td>0 1 1 1 1 0 0</td>
<td>1 0</td>
<td>0</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>1 1 1 1 1 1</td>
<td>0 1 0 1 0 1 1 1</td>
<td>0 1</td>
<td>1</td>
</tr>
<tr>
<td>0 0 1 1 0 1</td>
<td>0 1 0 0 1 0 1 1</td>
<td>0 1</td>
<td>1</td>
</tr>
</tbody>
</table>

Predict $X_s$ based on $X_{\setminus s} := \{X_s, \, t \neq s\}$. 
Graph selection via neighborhood regression

For each node \( s \in V \), compute (regularized) max. likelihood estimate:

\[
\hat{\theta}[s] := \arg \min_{\theta \in \mathbb{R}^{p-1}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; X_i \setminus s) + \lambda_n \|\theta\|_1 \right\}
\]

local log. likelihood
regularization

Predict \( X_s \) based on \( X_{\setminus s} := \{X_t, t \neq s\} \).
Graph selection via neighborhood regression

1. For each node $s \in V$, compute (regularized) max. likelihood estimate:

$$\hat{\theta}[s] := \arg \min_{\theta \in \mathbb{R}^{p-1}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\theta; X_i \setminus s) + \lambda_n \|\theta\|_1 \right\}$$

   - local log. likelihood
   - regularization

2. Estimate the local neighborhood $\hat{N}(s)$ as support of regression vector $\hat{\theta}[s] \in \mathbb{R}^{p-1}$.

Predict $X_s$ based on $X \setminus s := \{X_t, t \neq s\}$.
US Senate network (2004–2006 voting)
1 Lecture 1 (Today): Basics of sparse recovery
   ▶ Sparse linear systems: $\ell_0/\ell_1$ equivalence
   ▶ Noisy case: Lasso, $\ell_2$-bounds and variable selection

2 Lecture 2 (Tuesday): A more general theory
   ▶ A range of structured regularizers
     ★ Group sparsity
     ★ Low-rank matrices and nuclear norm regularization
     ★ Matrix decomposition and robust PCA
   ▶ Ingredients of a general understanding

3 Lecture 3 (Wednesday): High-dimensional kernel methods
   ▶ Curse-of-dimensionality for non-parametric regression
   ▶ Reproducing kernel Hilbert spaces
   ▶ A simple but optimal estimator
Noiseless linear models and basis pursuit

- under-determined linear system: unidentifiable without constraints
- say $\theta^* \in \mathbb{R}^p$ is sparse: supported on $S \subset \{1, 2, \ldots, p\}$.

\[ \ell_0\text{-optimization} \]
\[
\theta^* = \arg \min_{\theta \in \mathbb{R}^p} \|\theta\|_0 \\
X \theta = y
\]

\[ \ell_1\text{-relaxation} \]
\[
\hat{\theta} \in \arg \min_{\theta \in \mathbb{R}^p} \|\theta\|_1 \\
X \theta = y
\]

- Computationally intractable NP-hard
- Linear program (easy to solve) Basis pursuit relaxation
Noiseless \( \ell_1 \) recovery: Unrescaled sample size

Probability of recovery versus sample size \( n \).
Noiseless $\ell_1$ recovery: Rescaled

Prob. exact recovery vs. sample size ($\mu = 0$)

Prob. of exact recovery versus rescaled sample size $\alpha := \frac{n}{s \log(p/s)}$.
Definition

For a fixed $S \subset \{1, 2, \ldots, p\}$, the matrix $X \in \mathbb{R}^{n \times p}$ satisfies the restricted nullspace property w.r.t. $S$, or RN($S$) for short, if

$$\left\{ \Delta \in \mathbb{R}^p \mid X\Delta = 0 \right\} \cap \left\{ \Delta \in \mathbb{R}^p \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1 \right\} = \{0\}.$$

(Donoho & Xu, 2001; Feuer & Nemirovski, 2003; Cohen et al, 2009)
**Restricted nullspace: necessary and sufficient**

### Definition

For a fixed $S \subset \{1, 2, \ldots, p\}$, the matrix $X \in \mathbb{R}^{n \times p}$ satisfies the restricted nullspace property w.r.t. $S$, or $\text{RN}(S)$ for short, if

$$\left\{ \Delta \in \mathbb{R}^p \mid X\Delta = 0 \right\} \cap \left\{ \Delta \in \mathbb{R}^p \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1 \right\} = \{0\}.$$

(Donoho & Xu, 2001; Feuer & Nemirovski, 2003; Cohen et al, 2009)

### Proposition

Basis pursuit $\ell_1$-relaxation is exact for all $S$-sparse vectors $\iff X$ satisfies $\text{RN}(S)$.
Restricted nullspace: necessary and sufficient

**Definition**

For a fixed $S \subset \{1, 2, \ldots, p\}$, the matrix $X \in \mathbb{R}^{n \times p}$ satisfies the restricted nullspace property w.r.t. $S$, or $\text{RN}(S)$ for short, if

$$
\left\{ \Delta \in \mathbb{R}^p \mid X\Delta = 0 \right\} \cap \left\{ \Delta \in \mathbb{R}^p \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1 \right\} = \{0\}.
$$

(Donoho & Xu, 2001; Feuer & Nemirovski, 2003; Cohen et al, 2009)

**Proof (sufficiency):**

1. Error vector $\hat{\Delta} = \theta^* - \hat{\theta}$ satisfies $X\hat{\Delta} = 0$, and hence $\hat{\Delta} \in \mathbb{N}(X)$.

2. Show that $\hat{\Delta} \in \mathbb{C}(S)$

   **Optimality of $\hat{\theta}$:** $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1 = \|\theta^*_S\|_1$.

   **Sparsity of $\theta^*$:** $\|\hat{\theta}\|_1 = \|\theta^* + \hat{\Delta}\|_1 = \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1$.

   **Triangle inequality:** $\|\theta^*_S + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \geq \|\theta^*_S\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1$.

3. Hence, $\hat{\Delta} \in \mathbb{N}(X) \cap \mathbb{C}(S)$, and $(\text{RN}) \implies \hat{\Delta} = 0$. 
Illustration of restricted nullspace property

Consider \( \theta^* = (0, 0, \theta_3^*) \), so that \( S = \{3\} \).

Error vector \( \hat{\Delta} = \hat{\theta} - \theta^* \) belongs to the set

\[
\mathbb{C}(S; 1) := \{(\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3 \mid |\Delta_1| + |\Delta_2| \leq |\Delta_3|\}.
\]
Some sufficient conditions

How to verify RN property for a given sparsity $s$?

1. **Elementwise incoherence condition** (Donoho & Xuo, 2001; Feuer & Nem., 2003)

\[
\max_{j,k=1,...,p} \left| \frac{X^T X}{n} - I_{p \times p} \right|_{jk} \leq \frac{\delta_1}{s}
\]

\[
\begin{array}{c}
\text{n} \\
\uparrow \\
x_1 \quad x_j \quad x_k \quad x_p \\
\downarrow \\
p
\end{array}
\]
Some sufficient conditions

How to verify RN property for a given sparsity $s$?

1. **Elementwise incoherence condition**  (Donoho & Xuo, 2001; Feuer & Nem., 2003)

$$\max_{j,k=1,\ldots,p} \left| \left( \frac{X^TX}{n} - I_{p\times p} \right)_{jk} \right| \leq \frac{\delta_1}{s}$$

$$\begin{bmatrix}
\vdots \\
x_1 \\
x_j \\
x_k \\
x_p \\
\vdots
\end{bmatrix}$$

2. **Restricted isometry, or submatrix incoherence**  (Candès & Tao, 2005)

$$\max_{|U|\leq 2s} \left\| \left( \frac{X^TX}{n} - I_{p\times p} \right)_{UU} \right\|_{op} \leq \delta_{2s}.$$
Some sufficient conditions

How to verify RN property for a given sparsity $s$?

1. **Elementwise incoherence condition**  
   (Donoho & Xuo, 2001; Feuer & Nem., 2003)
   
   $\max_{j,k=1,...,p} \left| \left( \frac{X^T X}{n} - I_{p \times p} \right)_{jk} \right| \leq \frac{\delta_1}{s}$

   Matrices with i.i.d. sub-Gaussian entries: holds w.h.p. for $n = \Omega(s^2 \log p)$

2. **Restricted isometry, or submatrix incoherence**  
   (Candes & Tao, 2005)
   
   $\max_{|U| \leq 2s} \left\| \left( \frac{X^T X}{n} - I_{p \times p} \right)_{UU} \right\|_{op} \leq \delta_{2s}$.
Some sufficient conditions

How to verify RN property for a given sparsity $s$?

1. **Elementwise incoherence condition** (Donoho & Xuo, 2001; Feuer & Nem., 2003)

   $$\max_{j, k=1, \ldots, p} \left| \left( \frac{X^T X}{n} - I_{p \times p} \right)_{jk} \right| \leq \frac{\delta_1}{s}$$

   Matrices with i.i.d. sub-Gaussian entries: holds w.h.p. for $n = \Omega(s^2 \log p)$

2. **Restricted isometry, or submatrix incoherence** (Candes & Tao, 2005)

   $$\max_{|U| \leq 2s} \left\| \left( \frac{X^T X}{n} - I_{p \times p} \right)_{UU} \right\|_{op} \leq \delta_{2s}.$$  

   Matrices with i.i.d. sub-Gaussian entries: holds w.h.p. for $n = \Omega(s \log \frac{p}{s})$
Violating matrix incoherence (elementwise/RIP)

**Important:**
Incoherence/RIP conditions imply RN, but are far from necessary.
Very easy to violate them.....
Violating matrix incoherence (elementwise/RIP)

Form random design matrix

\[ X = \begin{bmatrix} x_1 & x_2 & \ldots & x_p \end{bmatrix} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad \text{each row } X_i \sim N(0, \Sigma), \text{ i.i.d.} \]

Example: For some \( \mu \in (0, 1) \), consider the covariance matrix

\[ \Sigma = (1 - \mu)I_{p \times p} + \mu 11^T. \]
Violating matrix incoherence (elementwise/RIP)

Form random design matrix

\[ X = \begin{bmatrix} x_1 & x_2 & \ldots & x_p \end{bmatrix} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad \text{each row } X_i \sim N(0, \Sigma), \text{ i.i.d.} \]

Example: For some \( \mu \in (0, 1) \), consider the covariance matrix

\[ \Sigma = (1 - \mu)I_{p \times p} + \mu 11^T. \]

- Elementwise incoherence violated: for any \( j \neq k \)

\[ \mathbb{P} \left[ \frac{\langle x_j, x_k \rangle}{n} \geq \mu - \epsilon \right] \geq 1 - c_1 \exp(-c_2 n \epsilon^2). \]
Violating matrix incoherence (elementwise/RIP)

Form random design matrix

\[ X = \begin{bmatrix} x_1 & x_2 & \ldots & x_p \end{bmatrix} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}, \text{ each row } X_i \sim N(0, \Sigma), \text{ i.i.d.} \]

Example: For some \( \mu \in (0, 1) \), consider the covariance matrix

\[ \Sigma = (1 - \mu)I_{p \times p} + \mu 11^T. \]

- **Elementwise incoherence violated:** for any \( j \neq k \)
  \[ \mathbb{P} \left[ \frac{\langle x_j, x_k \rangle}{n} \geq \mu - \epsilon \right] \geq 1 - c_1 \exp(-c_2 n \epsilon^2). \]

- **RIP constants tend to infinity** as \((n, |S|)\) increases:
  \[ \mathbb{P} \left[ \left\| \frac{X_S^T X_S}{n} - I_{s \times s} \right\|_2 \geq \mu (s - 1) - 1 - \epsilon \right] \geq 1 - c_1 \exp(-c_2 n \epsilon^2). \]
Noiseless $\ell_1$ recovery for $\mu = 0.5$

Probab. versus rescaled sample size $\alpha := \frac{n}{s \log(p/s)}$. 

Prob. exact recovery vs. sample size ($\mu = 0.5$)
Direct result for restricted nullspace/eigenvalues

**Theorem (Raskutti, W., & Yu, 2010)**

Consider a random design $X \in \mathbb{R}^{n \times p}$ with each row $X_i \sim N(0, \Sigma)$ i.i.d., and define $\kappa(\Sigma) = \max_{j=1,2,...,p} \Sigma_{jj}$. Then for universal constants $c_1, c_2$,

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{1}{2} \|\Sigma^{1/2}\theta\|_2 - 9\kappa(\Sigma) \sqrt{\frac{\log p}{n}} \|\theta\|_1 \quad \text{for all } \theta \in \mathbb{R}^p$$

with probability greater than $1 - c_1 \exp(-c_2 n)$. 

Direct result for restricted nullspace/eigenvalues

**Theorem (Raskutti, W., & Yu, 2010)**

Consider a random design $X \in \mathbb{R}^{n \times p}$ with each row $X_i \sim N(0, \Sigma)$ i.i.d., and define $\kappa(\Sigma) = \max_{j=1,2,...,p} \Sigma_{jj}$. Then for universal constants $c_1, c_2$,

$$\frac{\|X\theta\|_2}{\sqrt{n}} \geq \frac{1}{2} \|\Sigma^{1/2}\theta\|_2 - 9\kappa(\Sigma) \sqrt{\log p \frac{\log p}{n}} \|\theta\|_1$$

for all $\theta \in \mathbb{R}^p$

with probability greater than $1 - c_1 \exp(-c_2n)$.

- much less restrictive than incoherence/RIP conditions
- many interesting matrix families are covered
  - Toeplitz dependency
  - constant $\mu$-correlation (previous example)
  - covariance matrix $\Sigma$ can even be degenerate
  - extensions to sub-Gaussian matrices (Rudelson & Zhou, 2012)
- related results hold for generalized linear models
Easy verification of restricted nullspace

- for any $\Delta \in \mathbb{C}(S)$, we have

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{Sc}\|_1 \leq 2\|\Delta_S\| \leq 2\sqrt{s} \|\Delta\|_2$$

- applying previous result:

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \left\{ \lambda_{min}(\sqrt{\Sigma}) - 18\kappa(\Sigma) \sqrt{\frac{s \log p}{n}} \right\} \|\Delta\|_2.$$

\(\gamma(\Sigma)\)
Easy verification of restricted nullspace

- for any $\Delta \in \mathbb{C}(S)$, we have
  $$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{Sc}\|_1 \leq 2\|\Delta_S\| \leq 2\sqrt{s} \|\Delta\|_2$$

- applying previous result:
  $$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \left\{ \lambda_{min}(\sqrt{\Sigma}) - 18\kappa(\Sigma) \sqrt{\frac{s \log p}{n}} \right\} \gamma(\Sigma) \|\Delta\|_2.$$ 

- have actually proven much more than restricted nullspace....
Easy verification of restricted nullspace

- for any $\Delta \in \mathbb{C}(S)$, we have
  \[ \|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{Sc}\|_1 \leq 2\|\Delta_S\| \leq 2\sqrt{s}\|\Delta\|_2 \]

- applying previous result:
  \[ \frac{\|X\Delta\|_2}{\sqrt{n}} \geq \left\{ \lambda_{min}(\sqrt{\Sigma}) - 18\kappa(\Sigma) \sqrt{\frac{s \log p}{n}} \right\} \gamma(\Sigma) \|\Delta\|_2. \]

- have actually proven much more than restricted nullspace....

**Definition**

A design matrix $X \in \mathbb{R}^{n \times p}$ satisfies the *restricted eigenvalue* (RE) condition over $S$ (denote RE$(S)$) with parameters $\alpha \geq 1$ and $\gamma > 0$ if

\[ \frac{\|X\Delta\|_2}{\sqrt{n}} \geq \gamma \|\Delta\|_2 \quad \text{for all} \ \Delta \in \mathbb{R}^p \ \text{such that} \ \|\Delta_{Sc}\|_1 \leq \alpha \|\Delta_S\|_1. \]

(van de Geer, 2007; Bickel, Ritov & Tsybakov, 2008)
Lasso and restricted eigenvalues

Turning to noisy observations...

\[
\begin{align*}
\boldsymbol{y} & = \boldsymbol{X} \theta^* + \boldsymbol{w} \\
n \times & \quad n \times p \\
S & \quad S^c
\end{align*}
\]

**Estimator:** Lasso program

\[
\hat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \| \boldsymbol{y} - \boldsymbol{X} \theta \|_2^2 + \lambda_n \| \theta \|_1 \right\}.
\]

**Goal:** Obtain bounds on \( \| \hat{\theta}_{\lambda_n} - \theta^* \|_2 \) that hold with high probability.
Lasso bounds: Four simple steps

Let’s analyze constrained version:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\theta \|_2^2 \quad \text{such that} \quad \|\theta\|_1 \leq R = \|\theta^*\|_1.$$
Lasso bounds: Four simple steps

Let’s analyze constrained version:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\theta \|_2^2 \quad \text{such that} \quad \|\theta\|_1 \leq R = \|\theta^*\|_1.$$ 

(1) By optimality of $\hat{\theta}$ and feasibility of $\theta^*$:

$$\frac{1}{2n} \| y - X\hat{\theta} \|_2^2 \leq \frac{1}{2n} \| y - X\theta^* \|_2^2.$$
Lasso bounds: Four simple steps

Let’s analyze constrained version:

\[
\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\theta \|_2^2 \quad \text{such that} \quad \|\theta\|_1 \leq R = \|\theta^*\|_1.
\]

(1) By optimality of \(\hat{\theta}\) and feasibility of \(\theta^*\):

\[
\frac{1}{2n} \| y - X\hat{\theta} \|_2^2 \leq \frac{1}{2n} \| y - X\theta^* \|_2^2.
\]

(2) Derive a basic inequality: re-arranging in terms of \(\Delta = \hat{\theta} - \theta^*\):

\[
\frac{1}{n} \| X\Delta \|_2^2 \leq \frac{2}{n} \langle \Delta, X^T w \rangle.
\]
Lasso bounds: Four simple steps

Let’s analyze constrained version:

\[ \min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\theta \|_2^2 \quad \text{such that } \| \theta \|_1 \leq R = \| \theta^* \|_1. \]

---

1. By optimality of \( \hat{\theta} \) and feasibility of \( \theta^* \):

\[ \frac{1}{2n} \| y - X\hat{\theta} \|_2^2 \leq \frac{1}{2n} \| y - X\theta^* \|_2^2. \]

2. Derive a basic inequality: re-arranging in terms of \( \hat{\Delta} = \hat{\theta} - \theta^* \):

\[ \frac{1}{n} \| X\hat{\Delta} \|_2^2 \leq \frac{2}{n} \langle \hat{\Delta}, X^Tw \rangle. \]

3. Restricted eigenvalue for LHS; Hölder’s inequality for RHS

\[ \gamma \| \hat{\Delta} \|_2^2 \leq \frac{1}{n} \| X\hat{\Delta} \|_2^2 \leq \frac{2}{n} \langle \hat{\Delta}, X^Tw \rangle \leq 2\| \hat{\Delta} \|_1 \left\| \frac{X^Tw}{n} \right\|_\infty. \]
Lasso bounds: Four simple steps

Let’s analyze constrained version:

\[
\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \| y - X\theta \|_2^2 \quad \text{such that} \quad \| \theta \|_1 \leq R = \| \theta^\ast \|_1.
\]

(1) By optimality of \( \hat{\theta} \) and feasibility of \( \theta^\ast \):

\[
\frac{1}{2n} \| y - X\hat{\theta} \|_2^2 \leq \frac{1}{2n} \| y - X\theta^\ast \|_2^2.
\]

(2) Derive a basic inequality: re-arranging in terms of \( \hat{\Delta} = \hat{\theta} - \theta^\ast \):

\[
\frac{1}{n} \| X\hat{\Delta} \|_2^2 \leq \frac{2}{n} \langle \hat{\Delta}, X^T w \rangle.
\]

(3) Restricted eigenvalue for LHS; Hölder’s inequality for RHS

\[
\gamma \| \hat{\Delta} \|_2^2 \leq \frac{1}{n} \| X\hat{\Delta} \|_2^2 \leq \frac{2}{n} \langle \hat{\Delta}, X^T w \rangle \leq 2 \| \hat{\Delta} \|_1 \left\| \frac{X^T w}{n} \right\|_\infty.
\]

(4) As before, \( \hat{\Delta} \in \mathbb{C}(S) \), so that \( \| \hat{\Delta} \|_1 \leq 2\sqrt{s}\| \hat{\Delta} \|_2 \), and hence

\[
\| \hat{\Delta} \|_2 \leq \frac{4}{\gamma} \sqrt{s} \left\| \frac{X^T w}{n} \right\|_\infty.
\]
Lasso error bounds for different models

**Proposition**

Suppose that
- vector $\theta^*$ has support $S$, with cardinality $s$, and
- design matrix $X$ satisfies $\text{RE}(S)$ with parameter $\gamma > 0$.

For constrained Lasso with $R = \|\theta^*\|_1$ or regularized Lasso with $\lambda_n = 2\|X^Tw/n\|_\infty$, any optimal solution $\hat{\theta}$ satisfies the bound

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{4\sqrt{s}}{\gamma} \left\| \frac{X^Tw}{n} \right\|_\infty.$$
Lasso error bounds for different models

Proposition

Suppose that
- the vector $\theta^\ast$ has support $S$, with cardinality $s$, and
- the design matrix $X$ satisfies RE($S$) with parameter $\gamma > 0$.

For constrained Lasso with $R = \|\theta^\ast\|_1$ or regularized Lasso with $\lambda_n = 2\|X^T w / n\|_\infty$, any optimal solution $\hat{\theta}$ satisfies the bound

$$
\|\hat{\theta} - \theta^\ast\|_2 \leq \frac{4\sqrt{s}}{\gamma} \left\| \frac{X^T w}{n} \right\|_\infty.
$$

- this is a deterministic result on the set of optimizers
- various corollaries for specific statistical models
Lasso error bounds for different models

**Proposition**

Suppose that
- vector $\theta^*$ has support $S$, with cardinality $s$, and
- design matrix $X$ satisfies $RE(S)$ with parameter $\gamma > 0$.

For constrained Lasso with $R = \|\theta^*\|_1$ or regularized Lasso with $\lambda_n = 2\|X^T w/n\|_\infty$, any optimal solution $\hat{\theta}$ satisfies the bound

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{4\sqrt{s}}{\sqrt{\gamma}} \|\frac{X^T w}{n}\|_\infty.$$ 

- this is a deterministic result on the set of optimizers
- various corollaries for specific statistical models
  - Compressed sensing: $X_{ij} \sim N(0, 1)$ and bounded noise $\|w\|_2 \leq \sigma\sqrt{n}$
  - Deterministic design: $X$ with bounded columns and $w_i \sim N(0, \sigma^2)$

$$\|\frac{X^T w}{n}\|_\infty \leq \sqrt{\frac{3\sigma^2 \log p}{n}} \quad \text{w.h.p.} \quad \Rightarrow \quad \|\hat{\theta} - \theta^*\|_2 \leq \frac{4\sigma}{\gamma(L)} \sqrt{\frac{s \log p}{n}}.$$
Look-ahead to Lecture 2: A more general theory

Recap: Thus far.....

- Derived error bounds for basis pursuit and Lasso ($\ell_1$-relaxation)
- Seen importance of restricted nullspace and restricted eigenvalues
The big picture:

Lots of other estimators with same basic form:

\[ \hat{\theta}_{\lambda_n} \in \arg \min_{\theta \in \Omega} \left\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \right\}. \]
Look-ahead to Lecture 2: A more general theory

The big picture:

Lots of other estimators with same basic form:

\[ \hat{\theta}_{\lambda_n} \in \arg \min_{\theta \in \Omega} \left\{ \mathcal{L}(\theta; Z^n_1) + \lambda_n \mathcal{R}(\theta) \right\}. \]

Estimate \hspace{2cm} Loss function \hspace{2cm} Regularizer

Past years have witnessed an explosion of results (compressed sensing, covariance estimation, block-sparsity, graphical models, matrix completion...)

Question:

Is there a common set of underlying principles?
Some papers (www.eecs.berkeley.edu/~wainwrig)


