A primer on high-dimensional statistics: Lecture 1

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Simons Institute Workshop, Bootcamp Tutorials

Introduction

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

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

Introduction

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

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

- 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 d and n may be large (possibly $d \gg n$)
  - need for high-dimensional theory that provides non-asymptotic results for (n, d)

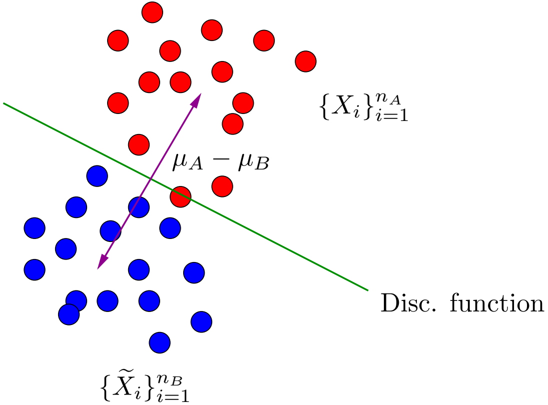
- curses and blessings of high dimensionality
  - exponential explosions in computational complexity
  - statistical curses (sample complexity)
  - concentration of measure

Key questions:

- What embedded low-dimensional structures are present in data?
- How can they can be exploited algorithmically?

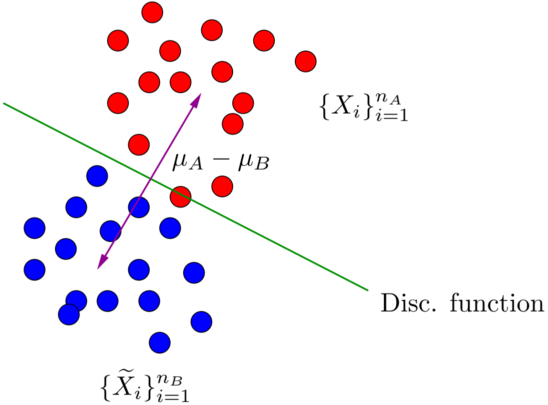
Vignette I: Linear discriminant analysis

Samples $\{X_1, \ldots, X_{n_A}\}$ from class A and $\{\widetilde{X}_1, \ldots, \widetilde{X}_{n_B}\}$ from class B



Vignette I: Linear discriminant analysis

Samples $\{X_1, \ldots, X_{n_A}\}$ from class A and $\{\widetilde{X}_1, \ldots, \widetilde{X}_{n_B}\}$ from class B



Optimal decision boundary in Gaussian case:

$$f(x) = \langle \mu_A - \mu_B, (\Sigma^{-1})\big(x - \frac{\mu_A + \mu_B}{2}\big)\rangle$$

with known shared variance Σ, and means μ_A, μ_B.

Classical vs. high-dimensional asymptotics

"Plug-in" principle: substitute estimates $\{\mu_A, \mu_B, \Sigma\}$ from given sample:

$$\widehat{f}(x) = \langle \widehat{\mu}_A - \widehat{\mu}_B,\, (\widehat{\Sigma})^{-1}\big(x - \frac{\widehat{\mu}_A + \widehat{\mu}_B}{2}\big)\rangle.$$

Classical analysis (say $\Sigma = I_{d \times d}$):

$$\mathbb{P}[\text{class. error}] \stackrel{n \rightarrow +\infty}{\longrightarrow} \underbrace{\Phi\big(\frac{-\|\mu_A - \mu_B\|_2}{2}\big)}_{\text{Tail function of standard normal}}$$

Classical vs. high-dimensional asymptotics

"Plug-in" principle: substitute estimates $\{\mu_A, \mu_B, \Sigma\}$ from given sample:

$$\widehat{f}(x) = \langle \widehat{\mu}_A - \widehat{\mu}_B,\, (\widehat{\Sigma})^{-1}\big(x - \frac{\widehat{\mu}_A + \widehat{\mu}_B}{2}\big)\rangle.$$

Classical analysis (say $\Sigma = I_{d \times d}$):

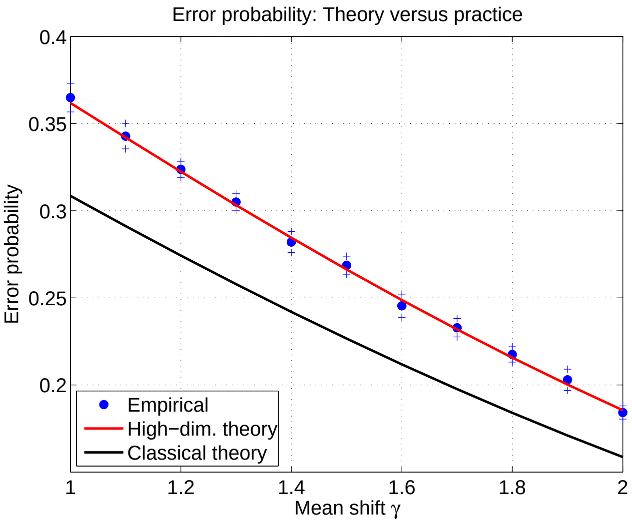
$$\mathbb{P}[\text{class. error}] \stackrel{n \to +\infty}{\longrightarrow} \underbrace{\Phi\big(\frac{-\|\mu_A - \mu_B\|_2}{2}\big)}_{\text{Tail function of standard normal}}$$

High-dimensional view: Kolmogorov, 1960s

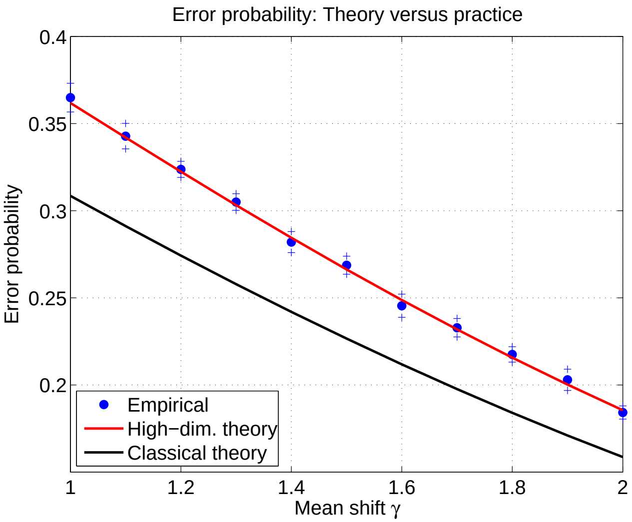
What happens if $(n_A, n_B, d) \to +\infty$ with

$$\frac{d}{n_A} \to \alpha, \quad \frac{d}{n_B} \to \alpha.$$

Error probability versus mean shift $\gamma = \|\mu_A - \mu_B\|_2$



Error probability versus mean shift $\gamma = \|\mu_A - \mu_B\|_2$



Kolmogorov prediction: $\Phi\left(-\frac{\gamma^2}{2\sqrt{\gamma^2+\alpha}}\right)$

Classical prediction: $\Phi\left(-\frac{\gamma}{2}\right)$.

Vignette II: Covariance estimation

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

Vignette II: Covariance estimation

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

Classical approach:
Estimate Σ via sample covariance matrix:

$$\widehat{\Sigma}_n := \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T}_{\text{average of } d \times d \text{ rank one matrices}}$$

Vignette II: Covariance estimation

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

Classical approach:
Estimate Σ via sample covariance matrix:

$$\widehat{\Sigma}_n := \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T}_{\text{average of } d \times d \text{ rank one matrices}}$$

Reasonable properties: (d fixed, n increasing)

- Unbiased: $\mathbb{E}[\widehat{\Sigma}_n] = \Sigma$
- Consistent: $\widehat{\Sigma}_n \xrightarrow{a.s.} \Sigma$ as $n \to +\infty$
- Asymptotic distributional properties available

Vignette II: Covariance estimation

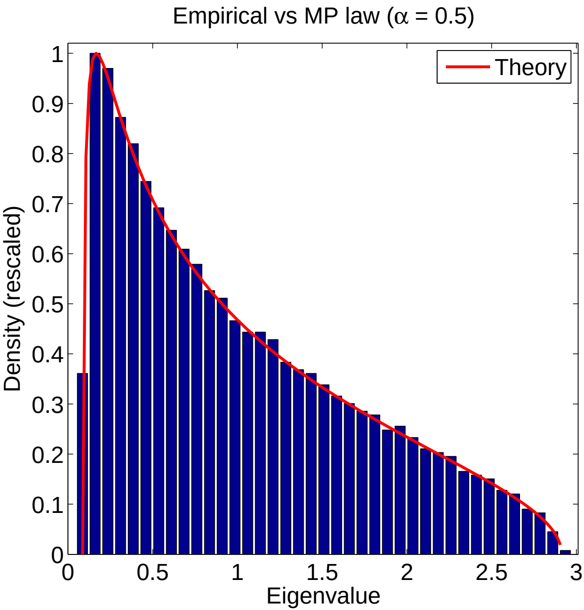
- want to estimate a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$
- given i.i.d. samples $X_i \sim N(0, \Sigma)$, for $i = 1, 2, \ldots, n$

Classical approach:
Estimate Σ via sample covariance matrix:

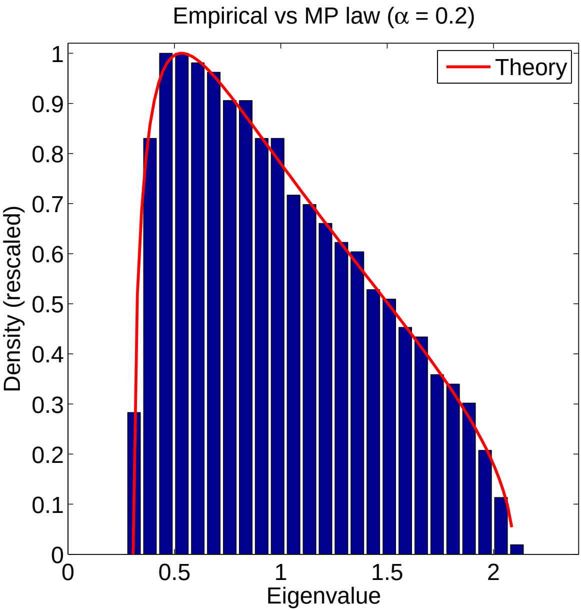
$$\widehat{\Sigma}_n := \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T}_{\text{average of } d \times d \text{ rank one matrices}}$$

An alternative experiment:

- Fix some $\alpha > 0$
- Study behavior over sequences with $\frac{d}{n} = \alpha$
- Does $\widehat{\Sigma}_{n(d)}$ converge to anything reasonable?

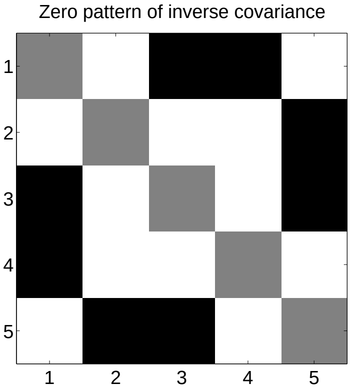


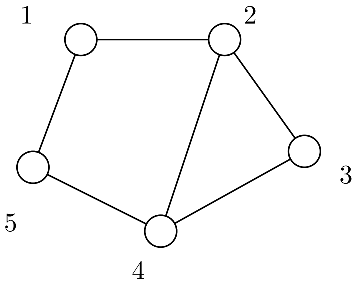
Marcenko & Pastur, 1967.



Marcenko & Pastur, 1967.

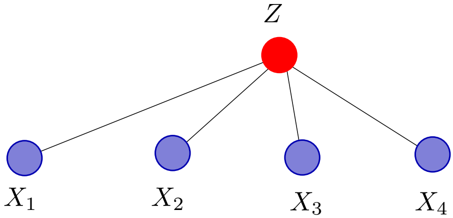
Low-dimensional structure: Gaussian graphical models





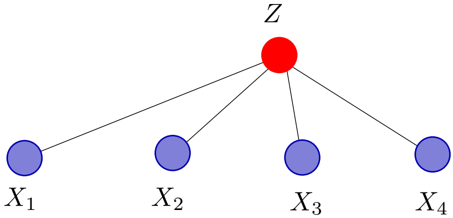
$$\mathbb{P}(x_1, x_2, \ldots, x_d) \propto \exp\big(-\frac{1}{2} x^T \Theta^* x\big).$$

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.

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:

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

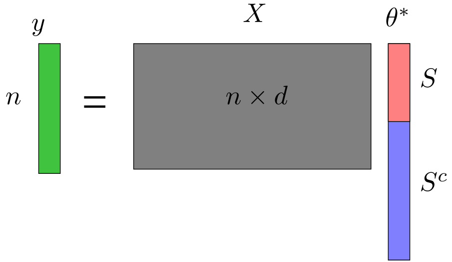
(Chandrasekaran, Parrilo & Willsky, 2010)

Outline

1. Lecture 1: Basics of sparse linear models
   - Sparse linear systems: ℓ_0/ℓ_1 equivalence
   - Noisy case: Lasso, ℓ_2-bounds and variable selection

2. Lecture 2: A more general theory
   - A range of structured regularizers
     - Group sparsity
     - Adaptive decompositions
     - Matrix completion and additive decomposition
     - Non-parametric problems
   - Ingredients of a general understanding

Noiseless linear models and basis pursuit



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

<u>ℓ_0-optimization</u>

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|_0 \quad X\theta = y$$

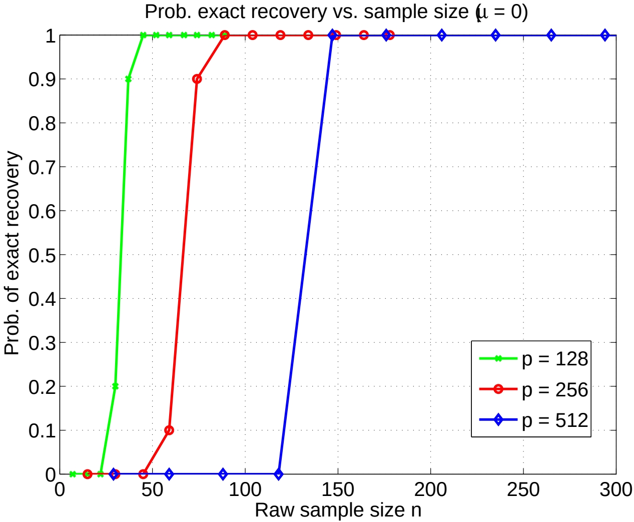
Computationally intractable
NP-hard

<u>ℓ_1-relaxation</u>

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad X\theta = y$$

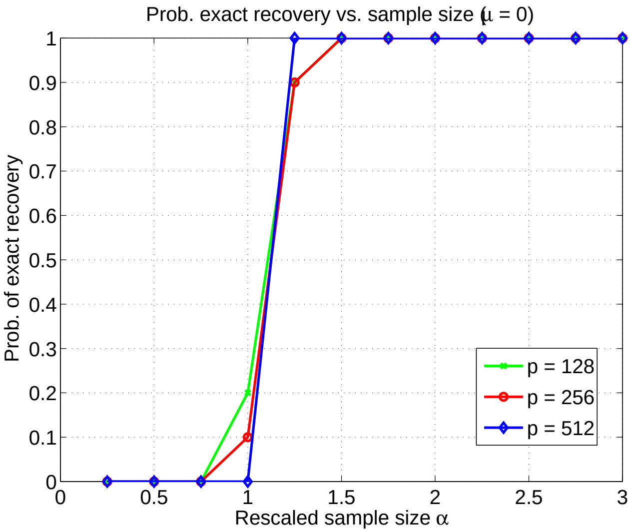
Linear program (easy to solve)
Basis pursuit relaxation

Noiseless ℓ_1 recovery: Unrescaled sample size



Probability of recovery versus sample size n.

Noiseless ℓ_1 recovery: Rescaled



Probabability of recovery versus rescaled sample size $\alpha := \frac{n}{s \log(d/s)}$.

Restricted nullspace: necessary and sufficient

Definition

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

$$\underbrace{\{\Delta \in \mathbb{R}^d \mid X\Delta = 0\}}_{\mathbb{N}(X)} \cap \underbrace{\{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}}_{\mathbb{C}(S)} = \{0\}.$$

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

Restricted nullspace: necessary and sufficient

Definition

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

$$\underbrace{\{\Delta \in \mathbb{R}^d \mid X\Delta = 0\}}_{\mathbb{N}(X)} \cap \underbrace{\{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}}_{\mathbb{C}(S)} = \{0\}.$$

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

Proposition

Basis pursuit ℓ_1-relaxation is exact for all S-sparse vectors $\Longleftrightarrow$ X satisfies $\mathrm{RN}(S)$.

Restricted nullspace: necessary and sufficient

Definition

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

$$\underbrace{\{\Delta \in \mathbb{R}^d \mid X\Delta = 0\}}_{\mathbb{N}(X)} \cap \underbrace{\{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}}_{\mathbb{C}(S)} = \{0\}.$$

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

Proof (sufficiency):

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

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

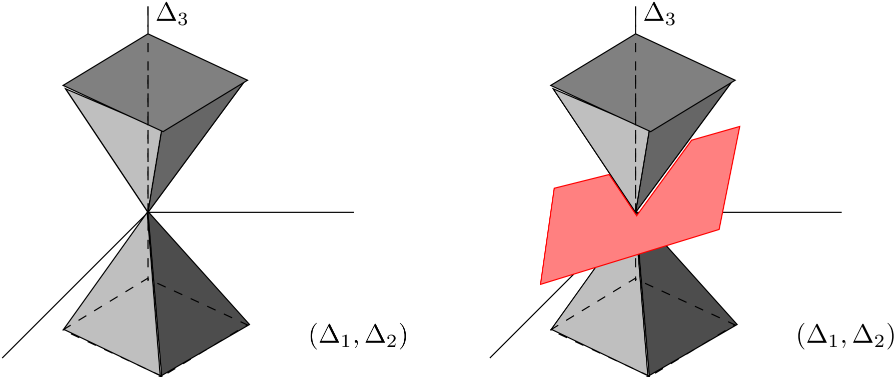
Optimality of $\widehat{\theta}$: $\|\widehat{\theta}\|_1 \leq \|\theta^*\|_1 = \|\theta^*_S\|_1$.

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

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

(3) Hence, $\widehat{\Delta} \in \mathbb{N}(X) \cap \mathbb{C}(S)$, and (RN) $\Longrightarrow \widehat{\Delta} = 0$.

Illustration of restricted nullspace property



- consider $\theta^* = (0, 0, \theta_3^*)$, so that $S = \{3\}$.
- error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ belongs to the set

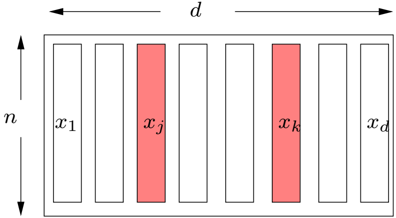
$$\mathbb{C}(S; 1) := \Big\{(\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3 \mid |\Delta_1| + |\Delta_2| \leq |\Delta_3|\Big\}.$$

Some sufficient conditions

How to verify RN property for a given sparsity s?

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

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

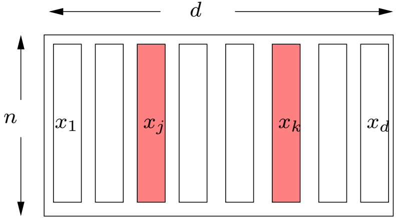


Some sufficient conditions

How to verify RN property for a given sparsity s?

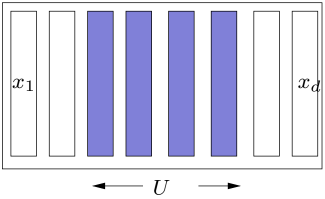
1. Elementwise incoherence condition (Donoho & Xuo, 2001; Feuer & Nem., 2003)

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



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

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

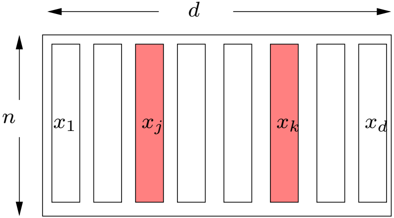


Some sufficient conditions

How to verify RN property for a given sparsity s?

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

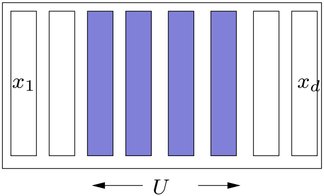
$$\max_{j,k=1,\ldots,d} \left| \left(\frac{X^T X}{n} - I_{d\times d} \right)_{jk} \right| \leq \frac{\delta_1}{s}$$



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

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

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

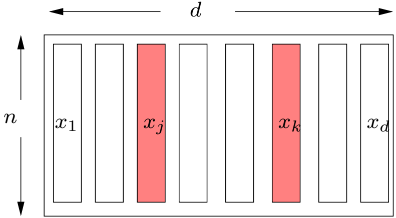


Some sufficient conditions

How to verify RN property for a given sparsity s?

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

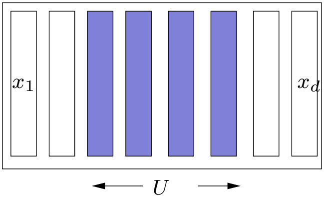
$$\max_{j,k=1,\ldots,d} \left| \left(\frac{X^T X}{n} - I_{d\times d} \right)_{jk} \right| \leq \frac{\delta_1}{s}$$



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

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

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



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

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

$$X = \underbrace{\begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix}}_{d \text{ columns}} = \underbrace{\begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}}_{n \text{ rows}} \in \mathbb{R}^{n \times d}, \qquad \text{each row } X_i \sim N(0, \Sigma), \text{ i.i.d.}$$

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

$$\Sigma = (1 - \mu) I_{d \times d} + \mu \mathbf{1}\mathbf{1}^T.$$

Violating matrix incoherence (elementwise/RIP)

Form random design matrix

$$X = \underbrace{\begin{bmatrix} x_1 & x_2 & \ldots & x_d \end{bmatrix}}_{d \text{ columns}} = \underbrace{\begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}}_{n \text{ rows}} \in \mathbb{R}^{n \times d}, \qquad \text{each row } X_i \sim N(0, \Sigma), \text{ i.i.d.}$$

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

$$\Sigma = (1 - \mu) I_{d \times d} + \mu \mathbf{1}\mathbf{1}^T.$$

- 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

$$X = \underbrace{\begin{bmatrix} x_1 & x_2 & \ldots & x_d \end{bmatrix}}_{d \text{ columns}} = \underbrace{\begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}}_{n \text{ rows}} \in \mathbb{R}^{n \times d}, \qquad \text{each row } X_i \sim N(0, \Sigma), \text{ i.i.d.}$$

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

$$\Sigma = (1 - \mu) I_{d \times d} + \mu \mathbf{1}\mathbf{1}^T.$$

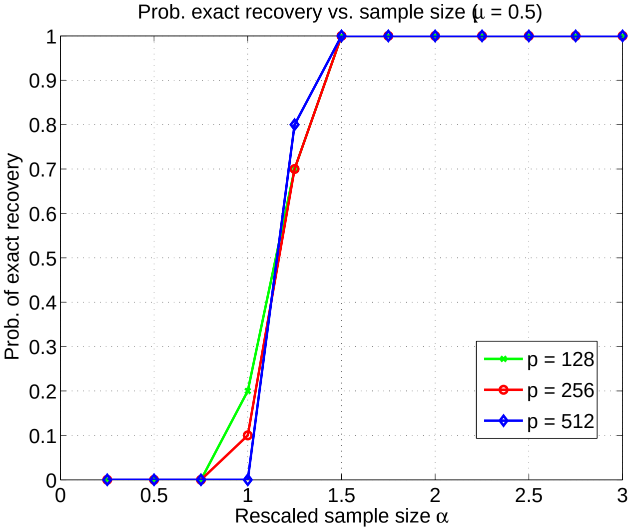
- Elementwise incoherence violated: for any $j \neq k$

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

- 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 ℓ_1 recovery for $\mu = 0.5$



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

Direct result for restricted nullspace/eigenvalues

Theorem (Raskutti, W., & Yu, 2010; Rudelson & Zhou, 2012)

Random Gaussian/sub-Gaussian matrix $X \in \mathbb{R}^{n \times d}$ *with i.i.d. rows, covariance* Σ*, and let* $\kappa^2 = \max_j \Sigma_{jj}$ *be the maximal variance. Then*

$$\frac{\|X\theta\|_2^2}{n} \geq c_1 \|\Sigma^{1/2}\theta\|_2^2 - c_2 \kappa^2(\Sigma) \, \frac{\log\left(e\, d\, (\frac{\|\theta\|_2}{\|\theta\|_1})^2\right)}{n} \, \|\theta\|_1^2 \qquad \textit{for all non-zero } \theta \in \mathbb{R}^d$$

with probability at least $1 - 2e^{-c_3 n}$.

Direct result for restricted nullspace/eigenvalues

Theorem (Raskutti, W., & Yu, 2010; Rudelson & Zhou, 2012)

Random Gaussian/sub-Gaussian matrix $X \in \mathbb{R}^{n \times d}$ *with i.i.d. rows, covariance* Σ*, and let* $\kappa^2 = \max_j \Sigma_{jj}$ *be the maximal variance. Then*

$$\frac{\|X\theta\|_2^2}{n} \geq c_1 \|\Sigma^{1/2}\theta\|_2^2 - c_2 \kappa^2(\Sigma) \, \frac{\log\left(e\, d\, (\frac{\|\theta\|_2}{\|\theta\|_1})^2\right)}{n} \, \|\theta\|_1^2 \qquad \text{for all non-zero } \theta \in \mathbb{R}^d$$

with probability at least $1 - 2e^{-c_3 n}$.

- many interesting matrix families are covered
  - Toeplitz dependency
  - constant μ-correlation (previous example)
  - covariance matrix Σ can even be degenerate
- related results hold for generalized linear models

Easy verification of restricted nullspace

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

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \;\leq\; 2\|\Delta_S\| \;\leq 2\sqrt{s}\,\|\Delta\|_2$$

- applying previous result:

$$\frac{\|X\Delta\|_2^2}{n} \geq \underbrace{\left\{c_1 \lambda_{min}(\Sigma) - 4c_2\kappa^2(\Sigma)\;\frac{s\log d}{n}\right\}}_{\gamma(\Sigma)}\;\|\Delta\|_2^2.$$

Easy verification of restricted nullspace

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

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \;\leq\; 2\|\Delta_S\| \;\leq 2\sqrt{s}\,\|\Delta\|_2$$

- applying previous result:

$$\frac{\|X\Delta\|_2^2}{n} \geq \underbrace{\left\{ c_1 \lambda_{min}(\Sigma) - 4 c_2 \kappa^2(\Sigma)\, \frac{s \log d}{n} \right\}}_{\gamma(\Sigma)} \|\Delta\|_2^2.$$

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

Easy verification of restricted nullspace

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

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \;\leq\; 2\|\Delta_S\| \;\leq 2\sqrt{s}\,\|\Delta\|_2$$

- applying previous result:

$$\frac{\|X\Delta\|_2^2}{n} \geq \underbrace{\left\{c_1 \lambda_{min}(\Sigma) - 4c_2 \kappa^2(\Sigma)\,\frac{s \log d}{n}\right\}}_{\gamma(\Sigma)} \|\Delta\|_2^2.$$

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

Definition

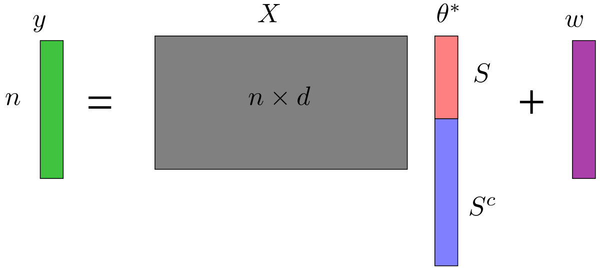
A design matrix $X \in \mathbb{R}^{n \times d}$ satisfies the *restricted eigenvalue* (RE) condition over S (denote RE(S)) with parameters $\alpha \geq 1$ and $\gamma > 0$ if

$$\frac{\|X\Delta\|_2^2}{n} \geq \gamma\,\|\Delta\|_2^2 \qquad \text{for all } \Delta \in \mathbb{R}^d \text{ such that } \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1.$$

(van de Geer, 2007; Bickel, Ritov & Tsybakov, 2008)

Lasso and restricted eigenvalues

Turning to noisy observations...



Estimator: Lasso program

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \Big\}.$$

Goal: Obtain bounds on { prediction error, parametric error, variable selection }.

Different error metrics

1. (In-sample) prediction error: $\|X(\widehat{\theta} - \theta^*)\|_2^2/n$
   - "weakest" error measure
   - appropriate when θ^* itself not of primary interest
   - strong dependence between columns of X possible (no RE needed)
   - proof technique: basic inequality

Different error metrics

1. (In-sample) prediction error: $\|X(\widehat{\theta} - \theta^*)\|_2^2/n$
   - "weakest" error measure
   - appropriate when θ^* itself not of primary interest
   - strong dependence between columns of X possible (no RE needed)
   - proof technique: basic inequality

2. parametric error: $\|\widehat{\theta} - \theta^*\|_r$ for some $r \in [1, \infty]$
   - appropriate for recovery problems
   - RE-type conditions appear in both lower/upper bounds
   - variable selection is not guaranteed
   - proof technique: basic inequality

Different error metrics

1. (In-sample) prediction error: $\|X(\widehat{\theta} - \theta^*)\|_2^2/n$
   - "weakest" error measure
   - appropriate when θ^* itself not of primary interest
   - strong dependence between columns of X possible (no RE needed)
   - proof technique: basic inequality

2. parametric error: $\|\widehat{\theta} - \theta^*\|_r$ for some $r \in [1, \infty]$
   - appropriate for recovery problems
   - RE-type conditions appear in both lower/upper bounds
   - variable selection is not guaranteed
   - proof technique: basic inequality

3. variable selection: is $\text{supp}(\widehat{\theta})$ equal to $\text{supp}(\theta^*)$?
   - appropriate when non-zero locations are of scientific interest
   - most stringent of all three criteria
   - requires incoherence or irrepresentability conditions on X
   - proof technique: primal-dual witness condition

Lasso ℓ_2-bounds: Four simple steps

Let's analyze constrained version:

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

Lasso ℓ_2-bounds: Four simple steps

Let's analyze constrained version:

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

---

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

$$\frac{1}{2n} \|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{2n} \|y - X\theta^*\|_2^2.$$

Lasso ℓ_2-bounds: Four simple steps

Let's analyze constrained version:

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

---

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

$$\frac{1}{2n} \|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{2n} \|y - X\theta^*\|_2^2.$$

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

$$\frac{1}{n} \|X\widehat{\Delta}\|_2^2 \leq \frac{2}{n} \langle \widehat{\Delta}, \, X^T w \rangle.$$

Lasso ℓ_2-bounds: Four simple steps

Let's analyze constrained version:

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

---

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

$$\frac{1}{2n} \|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{2n} \|y - X\theta^*\|_2^2.$$

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

$$\frac{1}{n} \|X\widehat{\Delta}\|_2^2 \leq \frac{2}{n} \langle \widehat{\Delta}, \, X^T w \rangle.$$

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

$$\gamma \|\widehat{\Delta}\|_2^2 \;\leq\; \frac{1}{n} \|X\widehat{\Delta}\|_2^2 \leq \frac{2}{n} \langle \widehat{\Delta}, \, X^T w \rangle \;\leq\; 2\|\widehat{\Delta}\|_1 \, \Big\| \frac{X^T w}{n} \Big\|_\infty.$$

Lasso ℓ_2-bounds: Four simple steps

Let's analyze constrained version:

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

---

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

$$\frac{1}{2n} \|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{2n} \|y - X\theta^*\|_2^2.$$

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

$$\frac{1}{n} \|X\widehat{\Delta}\|_2^2 \leq \frac{2}{n} \langle \widehat{\Delta}, \, X^T w \rangle.$$

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

$$\gamma \|\widehat{\Delta}\|_2^2 \;\leq\; \frac{1}{n} \|X\widehat{\Delta}\|_2^2 \leq \frac{2}{n} \langle \widehat{\Delta}, \, X^T w \rangle \;\leq\; 2\|\widehat{\Delta}\|_1 \, \Big\|\frac{X^T w}{n}\Big\|_\infty.$$

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

$$\|\widehat{\Delta}\|_2 \leq \frac{4}{\gamma} \, \sqrt{s} \, \Big\|\frac{X^T w}{n}\Big\|_\infty.$$

Lasso error bounds for different models

Proposition

Suppose that

- vector θ^* has support S, with cardinality s, and
- design matrix X satisfies $\mathrm{RE}(S)$ 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 $\widehat{\theta}$ satisfies the bound

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

Lasso error bounds for different models

Proposition

Suppose that

- vector θ^* has support S, with cardinality s, and
- design matrix X satisfies $\mathrm{RE}(S)$ 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 $\widehat{\theta}$ satisfies the bound

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

- 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 θ^* 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 $\widehat{\theta}$ satisfies the bound

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

- 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 $w_i \sim N(0, \sigma^2)$

$$\| \frac{X^T w}{n} \|_\infty \leq \sqrt{\frac{3\sigma^2 \log d}{n}} \quad \text{w.h.p.} \implies \|\widehat{\theta} - \theta^*\|_2 \leq \frac{4\sigma}{\gamma} \sqrt{3 \, \frac{s \log d}{n}}.$$

Extension to an oracle inequality

Previous theory assumed that θ^* was "hard" sparse. Not realistic in practice.

Extension to an oracle inequality

Previous theory assumed that θ^* was "hard" sparse. Not realistic in practice.

Theorem (An oracle inequality)

Suppose that least-squares loss satisfies γ-RE condition. Then for $\lambda_n \geq \max\{2\|\frac{X^T w}{n}\|_\infty, \sqrt{\frac{\log d}{n}}\}$, any optimal Lasso solution satisfies

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \min_{S \subseteq \{1,\ldots,d\}} \Big\{ \underbrace{\frac{9}{4}\frac{\lambda_n^2}{\gamma^2}|S|}_{\text{estimation error}} + \underbrace{\frac{2\lambda_n}{\gamma}\|\theta^*_{S^c}\|_1}_{\text{approximation error}} \Big\}.$$

(cf. Bunea et al., 2007; Buhlmann and van de Geer, 2009; Koltchinski et al., 2011)

Extension to an oracle inequality

Previous theory assumed that θ^* was "hard" sparse. Not realistic in practice.

Theorem (An oracle inequality)

Suppose that least-squares loss satisfies γ-RE condition. Then for $\lambda_n \geq \max\{2\|\frac{X^T w}{n}\|_\infty, \sqrt{\frac{\log d}{n}}\}$, any optimal Lasso solution satisfies

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \min_{S \subseteq \{1,\ldots,d\}} \Big\{ \underbrace{\frac{9}{4}\frac{\lambda_n^2}{\gamma^2}|S|}_{\text{estimation error}} + \underbrace{\frac{2\lambda_n}{\gamma}\|\theta^*_{S^c}\|_1}_{\text{approximation error}} \Big\}.$$

- when θ^* is exactly sparse, set $S = \text{supp}(\theta^*)$ to recover previous result

(cf. Bunea et al., 2007; Buhlmann and van de Geer, 2009; Koltchinski et al., 2011)

Extension to an oracle inequality

Previous theory assumed that θ^* was "hard" sparse. Not realistic in practice.

Theorem (An oracle inequality)

Suppose that least-squares loss satisfies γ-RE condition. Then for $\lambda_n \geq \max\{2\|\frac{X^T w}{n}\|_\infty, \sqrt{\frac{\log d}{n}}\}$, *any optimal Lasso solution satisfies*

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \min_{S \subseteq \{1,\ldots,d\}} \Big\{ \underbrace{\frac{9}{4}\frac{\lambda_n^2}{\gamma^2}|S|}_{\text{estimation error}} + \underbrace{\frac{2\lambda_n}{\gamma}\|\theta^*_{S^c}\|_1}_{\text{approximation error}} \Big\}.$$

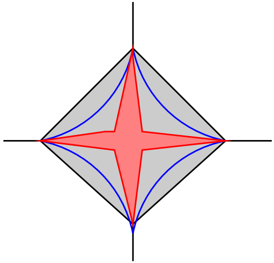
- when θ^* is exactly sparse, set $S = \text{supp}(\theta^*)$ to recover previous result
- more generally, choose S adaptively to trade-off estimation error versus approximation error

(cf. Bunea et al., 2007; Buhlmann and van de Geer, 2009; Koltchinski et al., 2011)

Consequences for ℓ_q-"ball" sparsity

- for some $q \in [0,1]$, say θ^* belongs to ℓ_q-"ball"

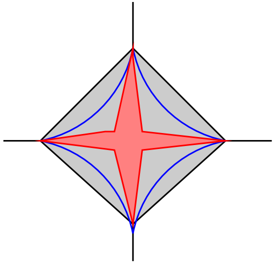
$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q\big\}.$$



Consequences for ℓ_q-"ball" sparsity

- for some $q \in [0, 1]$, say θ^* belongs to ℓ_q-"ball"

$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q\big\}.$$



Corollary

Consider the linear model $y = X\theta^* + w$, where X satisfies lower RE conditions, and w has i.i.d σ sub-Gaussian entries. For $\theta^* \in \mathbb{B}_q(R_q)$, any Lasso solution satisfies (w.h.p.)

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim R_q \left(\frac{\sigma^2 \log d}{n}\right)^{1-q/2}.$$

Are these good results? Minimax theory

- let $\mathcal{P}$ be a family of probability distributions
- consider a parameter $\mathbb{P} \mapsto \theta(\mathbb{P})$
- define a metric ρ on the parameter space

Are these good results? Minimax theory

- let $\mathcal{P}$ be a family of probability distributions
- consider a parameter $\mathbb{P} \mapsto \theta(\mathbb{P})$
- define a metric ρ on the parameter space

Definition (Minimax rate)

The minimax rate for $\theta(\mathcal{P})$ with metric ρ is given

$$\mathfrak{M}_n(\theta(\mathcal{P}); \rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\big[\rho^2(\widehat{\theta}, \theta(\mathbb{P}))\big],$$

where the infimum ranges over all measureable functions of n samples.

Are these good results? Minimax theory

Definition (Minimax rate)

The minimax rate for $\theta(\mathcal{P})$ with metric ρ is given

$$\mathfrak{M}_n(\theta(\mathcal{P}); \rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}\big[\rho^2(\widehat{\theta}, \theta(\mathbb{P}))\big],$$

where the infimum ranges over all measureable functions of n samples.

Concrete example:

- let $\mathcal{P}$ be family of sparse linear regression problems with $\theta^* \in \mathbb{B}_q(R_q)$
- consider ℓ_2-error metric $\rho^2(\widehat{\theta}, \theta) = \|\widehat{\theta} - \theta\|_2^2$

Are these good results? Minimax theory

Definition (Minimax rate)

The minimax rate for $\theta(\mathcal{P})$ with metric ρ is given

$$\mathfrak{M}_n(\theta(\mathcal{P});\rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\big[\rho^2(\widehat{\theta}, \theta(\mathbb{P}))\big],$$

where the infimum ranges over all measureable functions of n samples.

Concrete example:

- let $\mathcal{P}$ be family of sparse linear regression problems with $\theta^* \in \mathbb{B}_q(R_q)$
- consider ℓ_2-error metric $\rho^2(\widehat{\theta}, \theta) = \|\widehat{\theta} - \theta\|_2^2$

Theorem (Raskutti, W. & Yu, 2011)

Under "mild" conditions on design X and radius R_q, we have

$$\mathfrak{M}_n\big(\mathbb{B}_q(R_q); \|\cdot\|_2\big) \asymp R_q \Big(\frac{\sigma^2 \, \log d}{n}\Big)^{1-\frac{q}{2}}.$$

see Donoho & Johnstone, 1994 for normal sequence model

Look-ahead to Lecture 2: A more general theory

Recap: Thus far.....

- Derived error bounds for basis pursuit and Lasso (ℓ_1-relaxation)
- Seen importance of restricted nullspace and restricted eigenvalues
- Touched upon notion of oracle inequality and minimax rates

Look-ahead to Lecture 2: A more general theory

The big picture:

Lots of other estimators with same basic form:

$$\underbrace{\widehat{\theta}_{\lambda_n}}_{\text{Estimate}} \in \arg\min_{\theta \in \Omega} \Big\{ \underbrace{\mathcal{L}(\theta; Z_1^n)}_{\text{Loss function}} + \lambda_n \underbrace{\mathcal{R}(\theta)}_{\text{Regularizer}} \Big\}.$$

Look-ahead to Lecture 2: A more general theory

The big picture:

Lots of other estimators with same basic form:

$$\underbrace{\widehat{\theta}_{\lambda_n}}_{\text{Estimate}} \in \arg\min_{\theta \in \Omega} \Big\{ \underbrace{\mathcal{L}(\theta; Z_1^n)}_{\text{Loss function}} + \lambda_n \underbrace{\mathcal{R}(\theta)}_{\text{Regularizer}} \Big\}.$$

Past years have witnessed an explosion of results (graph estimation, matrix completion, matrix decomposition, nonparametric regression...)

Question:

Is there a common set of underlying principles?