High-dimensional statistics: Some progress and challenges ahead

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

University College, London Master Class: Lecture 1

Joint work with: Alekh Agarwal, Arash Amini, Po-Ling Loh, Sahand Negahban, Garvesh Raskutti, Pradeep Ravikumar, Bin Yu.

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 can 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 $i = 1, 2, \ldots, n$

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 Σ via sample covariance matrix:

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

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 Σ via sample covariance matrix:

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

Reasonable properties: (p 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 I: High-dimensional matrix estimation

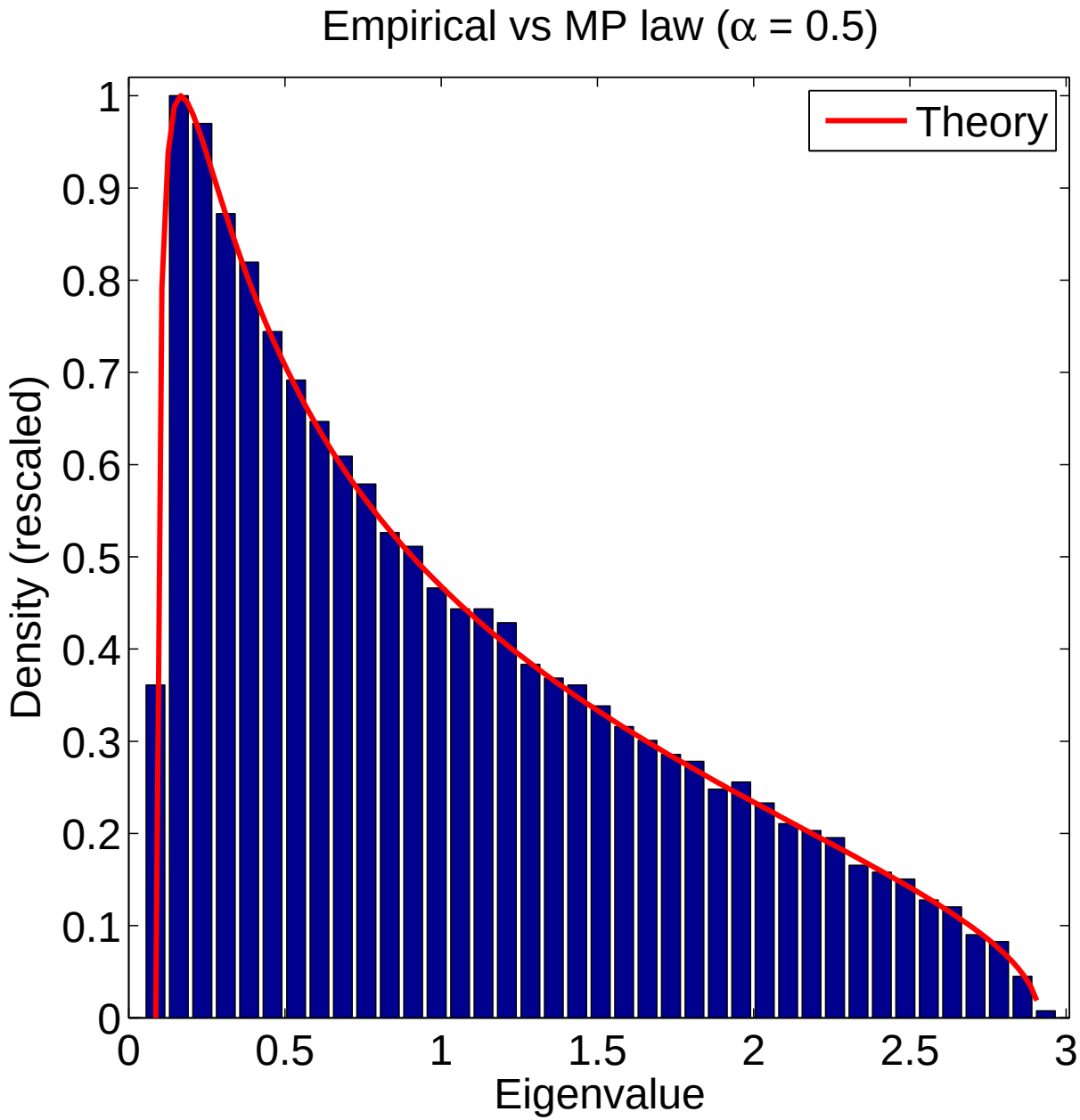
- want to estimate a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$
- given i.i.d. samples $X_i \sim N(0, \Sigma)$, for $i = 1, 2, \ldots, n$

Classical approach:
Estimate Σ via sample covariance matrix:

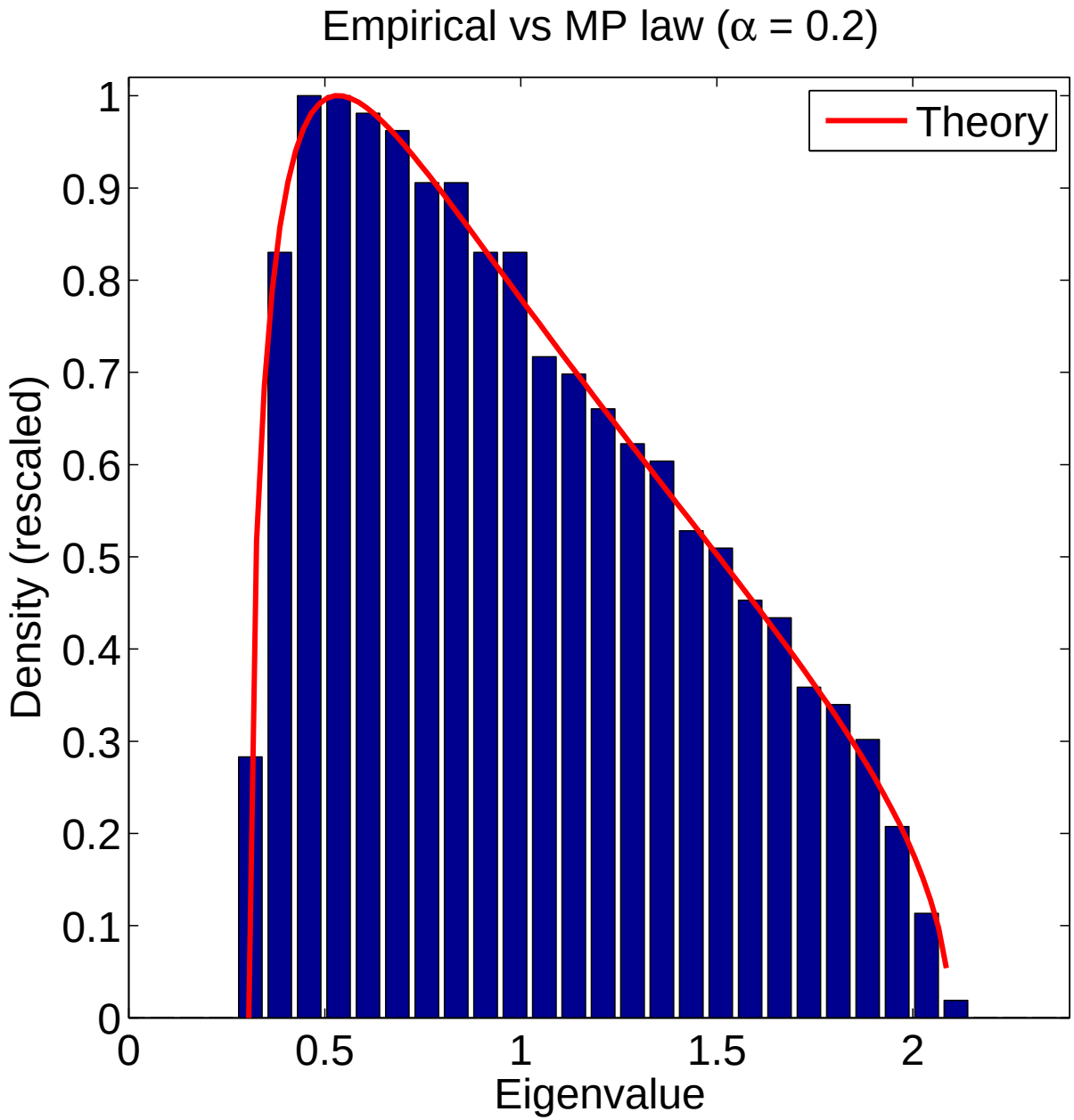
$$\widehat{\Sigma}_n := \underbrace{\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T}_{\text{average of } p \times p \text{ rank one matrices}}$$

An alternative experiment:

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

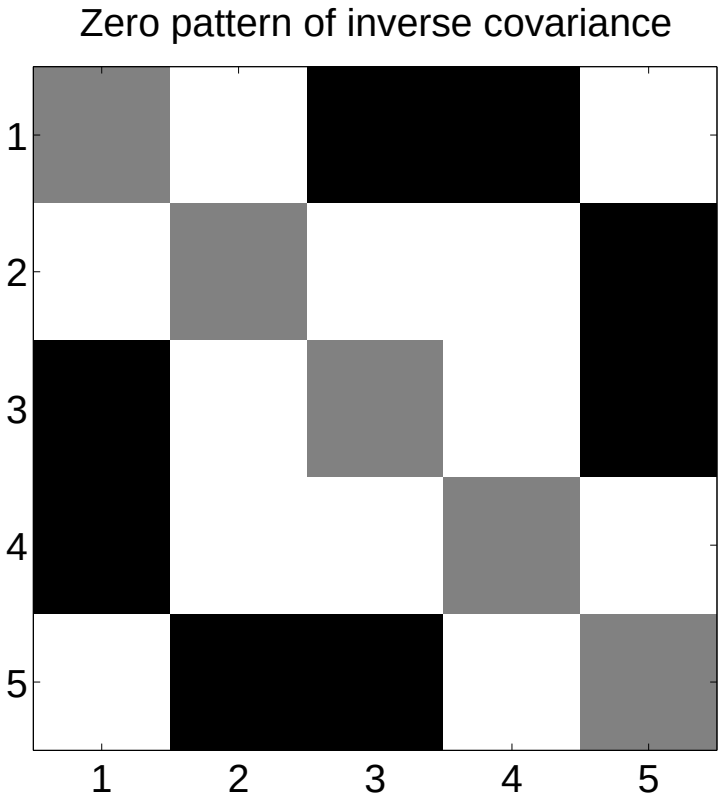


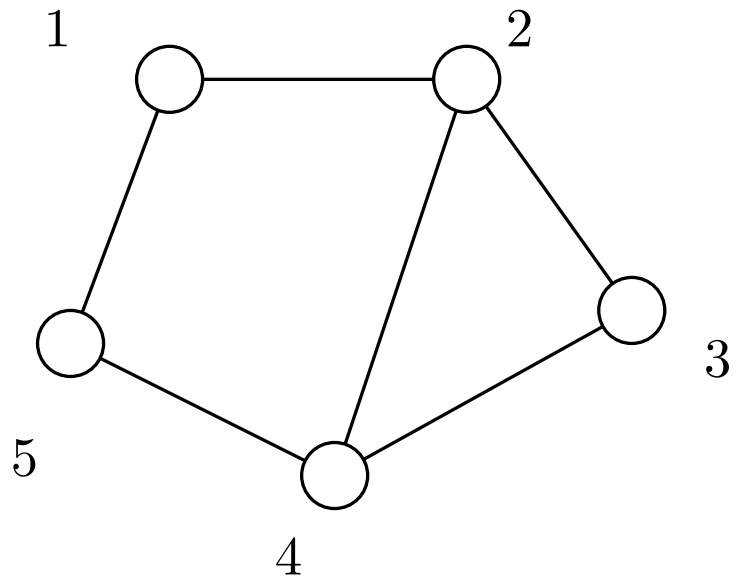
Marcenko & Pastur, 1967.



Marcenko & Pastur, 1967.

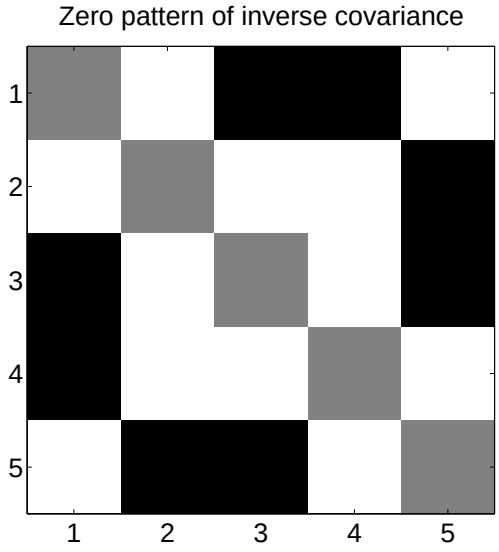
Low-dimensional structure: Gaussian graphical models

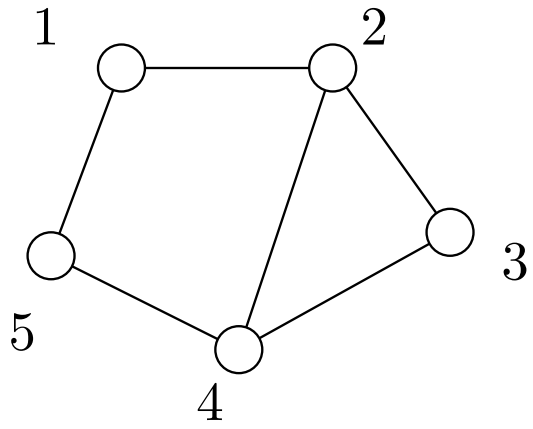




$$\mathbb{P}(x_1, x_2, \ldots, x_p) \propto \exp\Big(-\frac{1}{2} x^T \Theta^* x \Big).$$

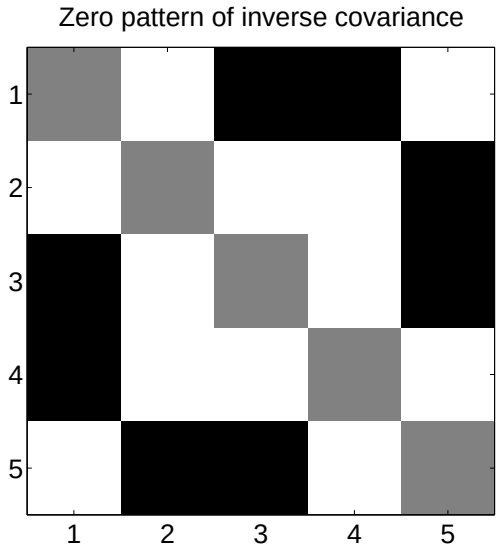
Maximum-likelihood with ℓ_1-regularization

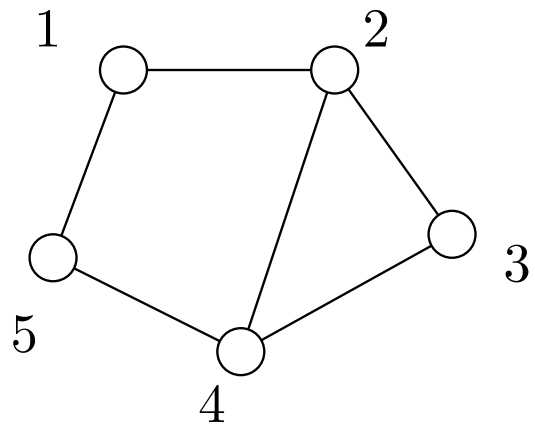




Set-up: Samples from random vector with sparse covariance Σ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{p \times p}$.

Maximum-likelihood with ℓ_1-regularization





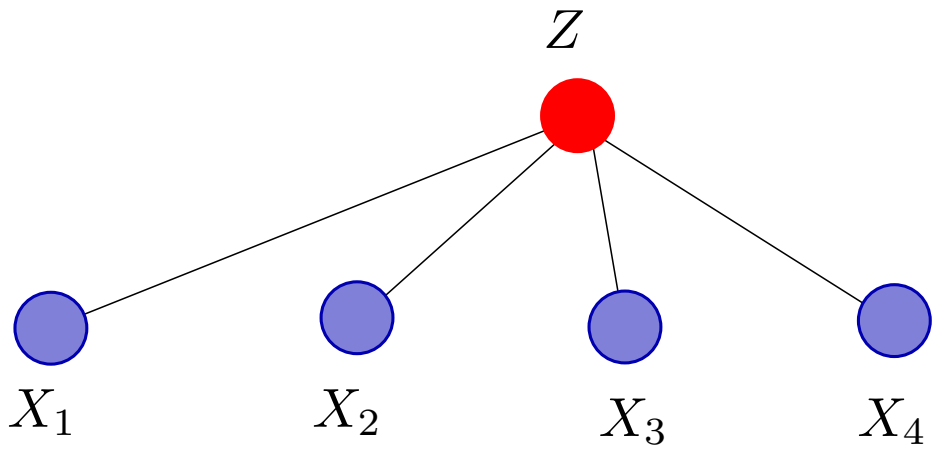
Set-up: Samples from random vector with sparse covariance Σ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{p \times p}$.

Estimator (for inverse covariance)

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

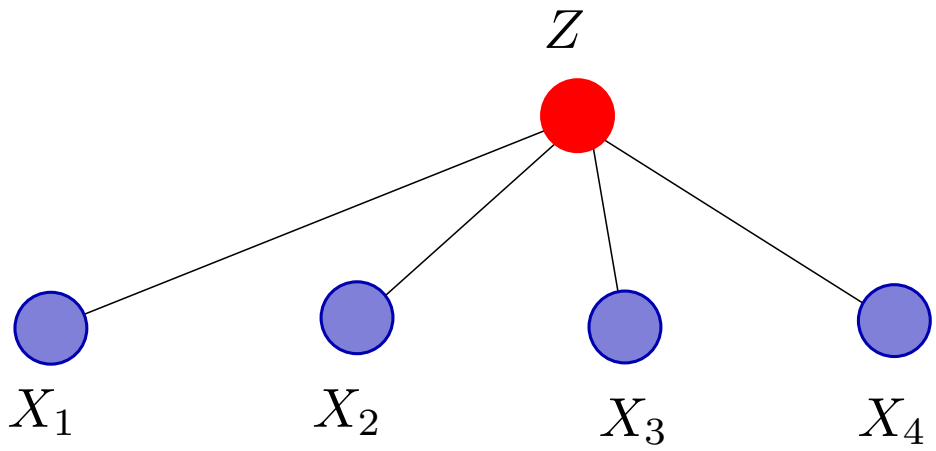
Some past work: Yuan & Lin, 2006; d'Asprémont et al., 2007; Bickel & Levina, 2007; El Karoui, 2007; d'Aspremont et al., 2007; Rothman et al., 2007; Zhou et al., 2007; Friedman et al., 2008; Lam & Fan, 2008; Ravikumar et al., 2008; Zhou, Cai & Huang, 2009

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



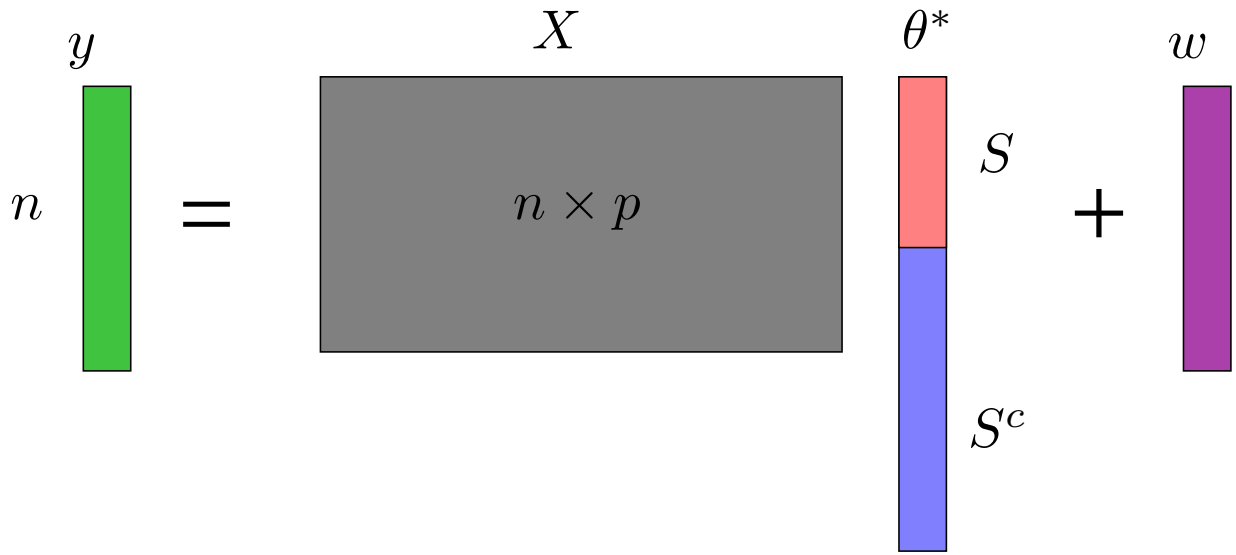
Problems with hidden variables: conditioned on hidden Z, vector $X = (X_1, X_2, X_3, X_4)$ is Gauss-Markov.

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)

Vignette II: High-dimensional sparse linear regression



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

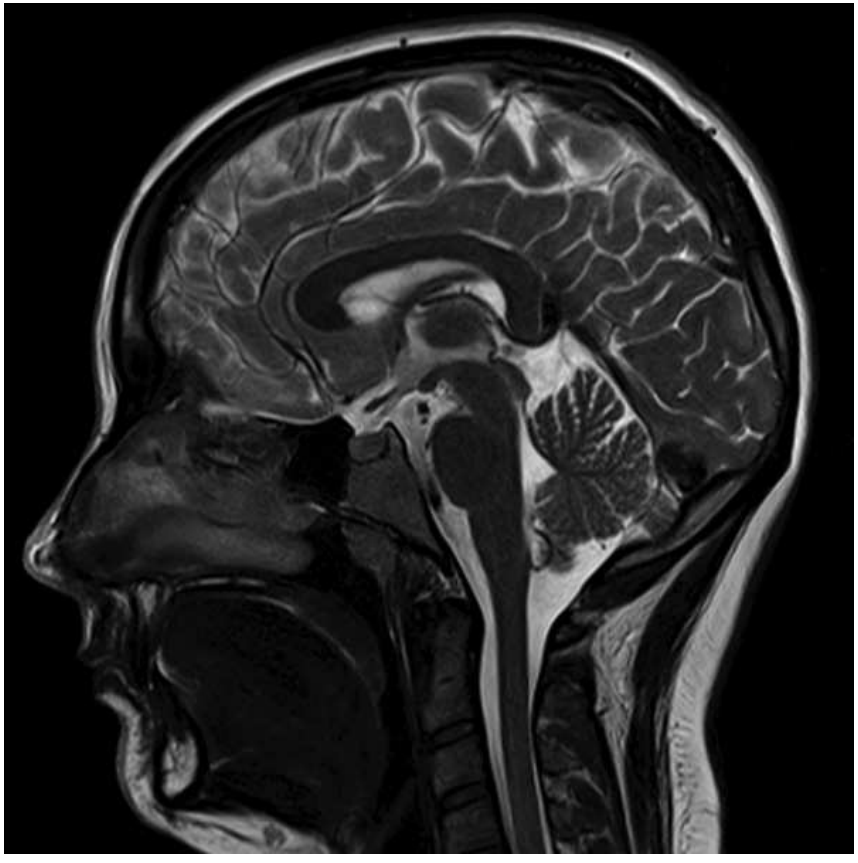
Estimator: Lasso program

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

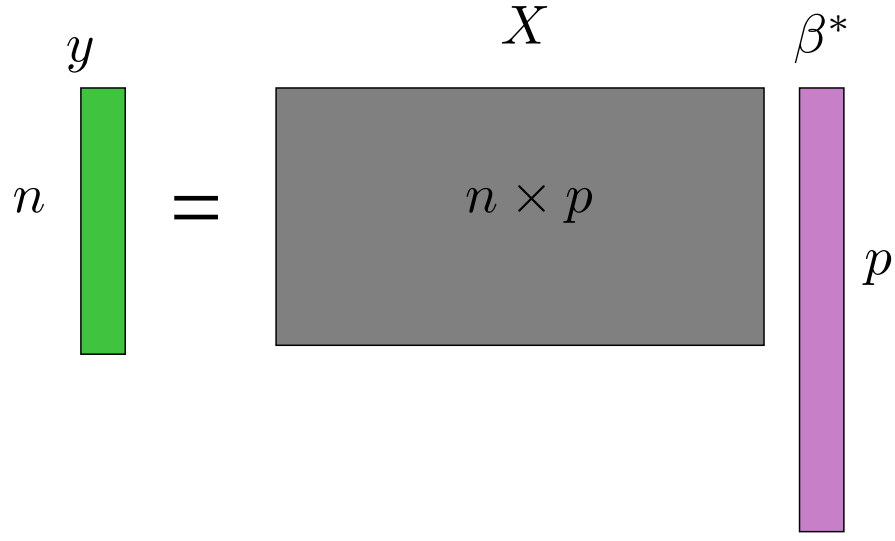
Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Efron et al., 2004; Meinshausen & Buhlmann, 2005; Candes & Tao, 2005; Donoho, 2005; Haupt & Nowak, 2005; Zhou & Yu, 2006; Zou, 2006; Koltchinskii, 2007; van

Application A: Compressed sensing

(Donoho, 2005; Candes & Tao, 2005)



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



(b) Compute n random projections

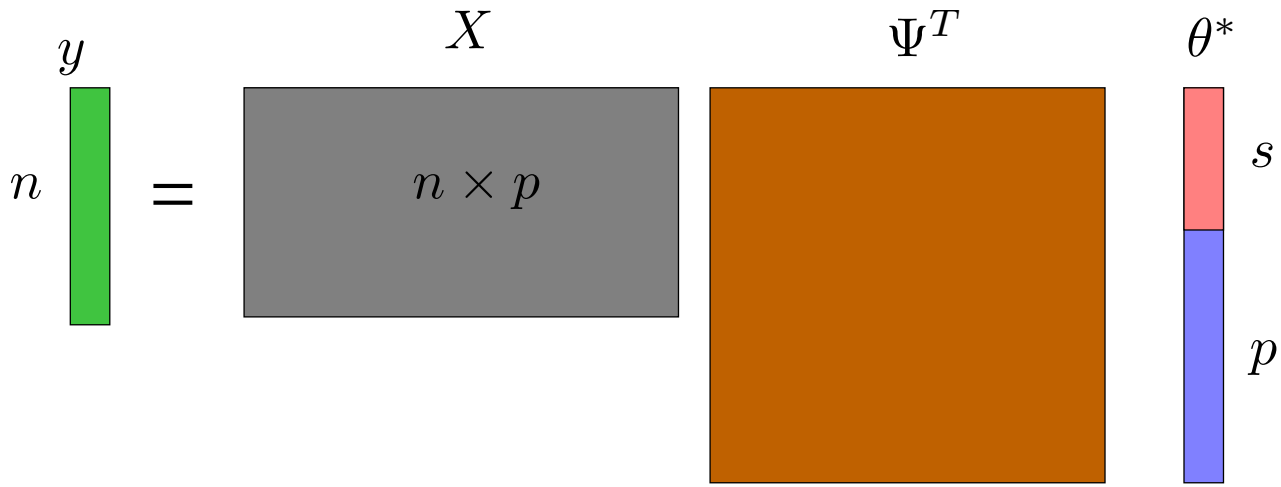
Application A: Compressed sensing

(Donoho, 2005; Candes & Tao, 2005)

In practice, signals are sparse in a transform domain:

$$\theta^* := \Psi\beta^* \quad \text{is a sparse signal,}$$

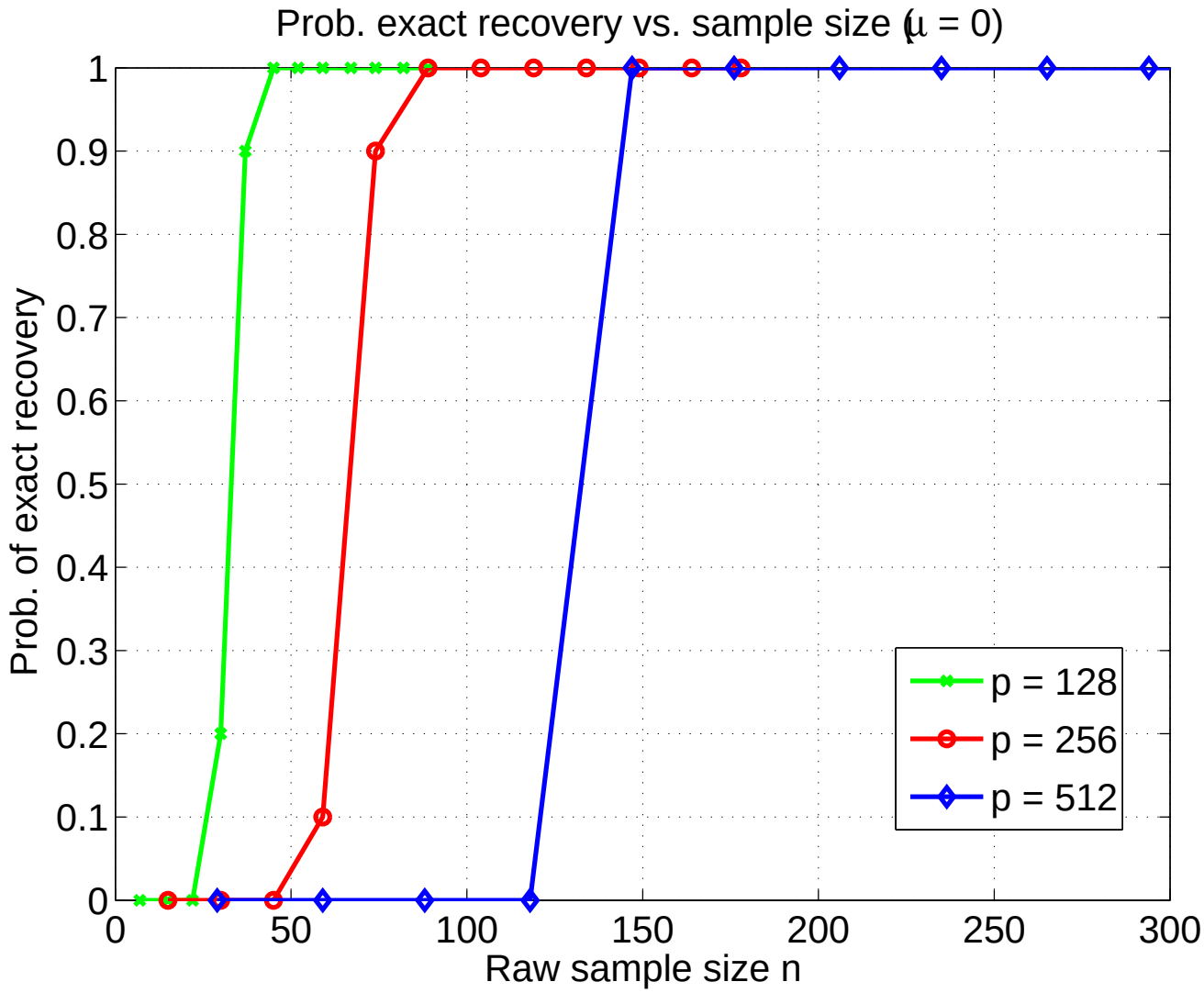
where Ψ is an orthonormal matrix.



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

$$y = \widetilde{X}\,\theta \qquad \text{where } \widetilde{X} = X\Psi^T \text{ is another random matrix.}$$

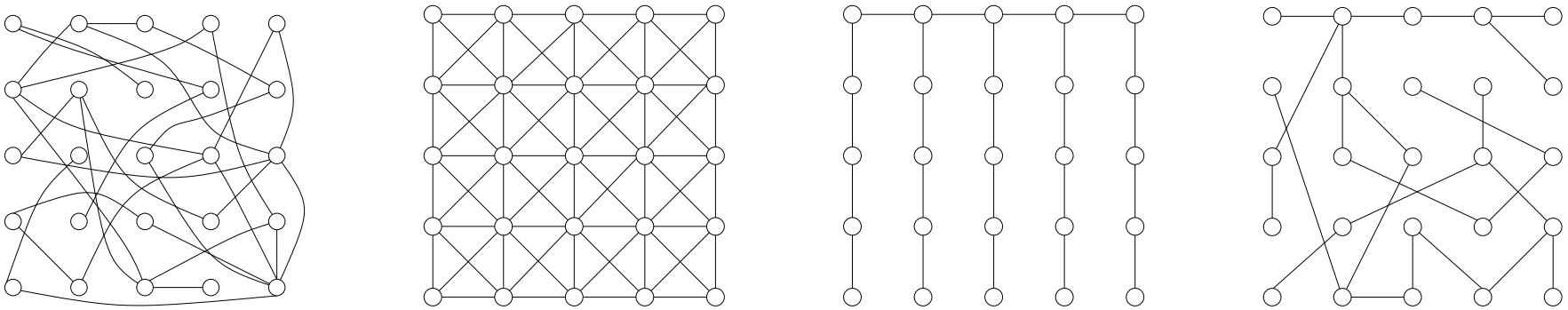
Noiseless ℓ_1 recovery: Unrescaled sample size



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:

$$\mathbb{P}(x_1, \ldots, x_p; \theta) \quad \propto \quad \exp\Big\{ \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\}.$$

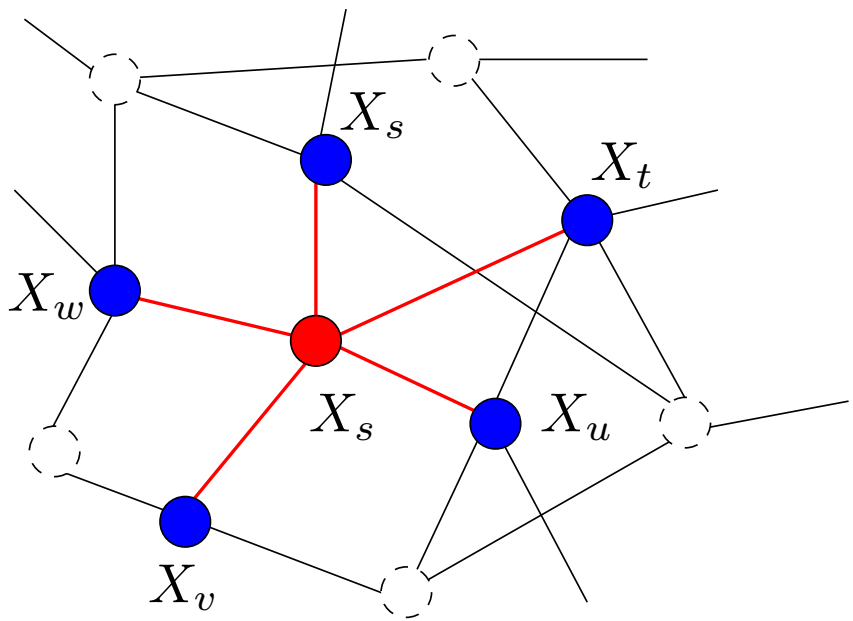
- given n independent and identically distributed (i.i.d.) samples of $X = (X_1, \ldots, X_p)$, identify the underlying graph structure

Pseudolikelihood and neighborhood regression

- Markov properties encode neighborhood structure:

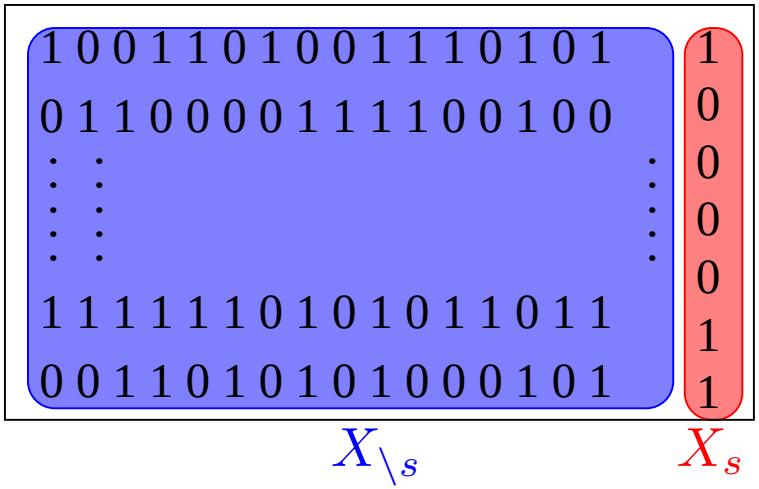
$$\underbrace{(X_s \quad | \quad X_{V\setminus s})}_{\text{Condition on full graph}} \stackrel{d}{=} \underbrace{(X_s \quad | \quad X_{N(s)})}_{\text{Condition on Markov blanket}}$$

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



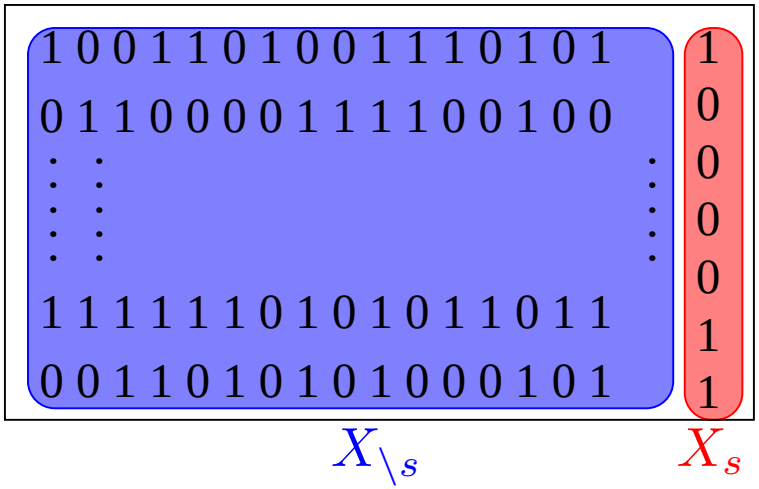
- basis of pseudolikelihood method (Besag, 1974)
- basis of many graph learning algorithm (Friedman et al., 1999; Csiszar & Talata, 2005; Abeel et al., 2006; Meinshausen & Buhlmann, 2006)

Graph selection via neighborhood regression



Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

Graph selection via neighborhood regression

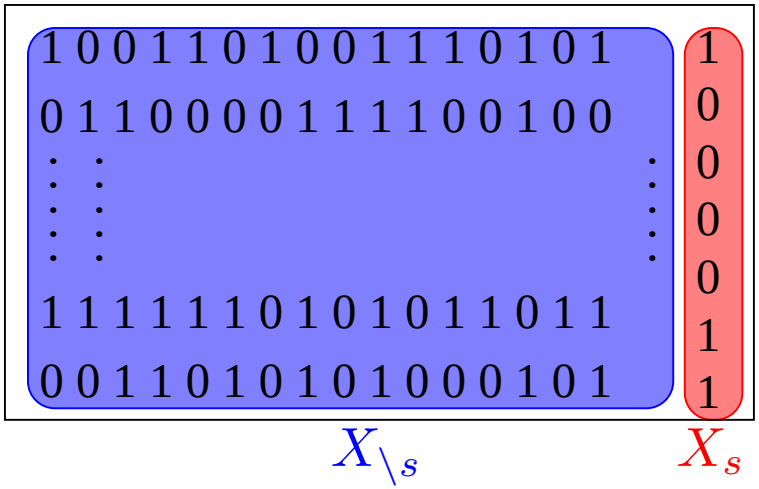


Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

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

$$\widehat{\theta}[s] \quad := \quad \arg\min_{\theta \in \mathbb{R}^{p-1}} \left\{ \quad -\frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathcal{L}(\theta; X_{i\,\setminus s})}_{\text{local log. likelihood}} \quad + \quad \lambda_n \underbrace{\|\theta\|_1}_{\text{regularization}} \right\}$$

Graph selection via neighborhood regression



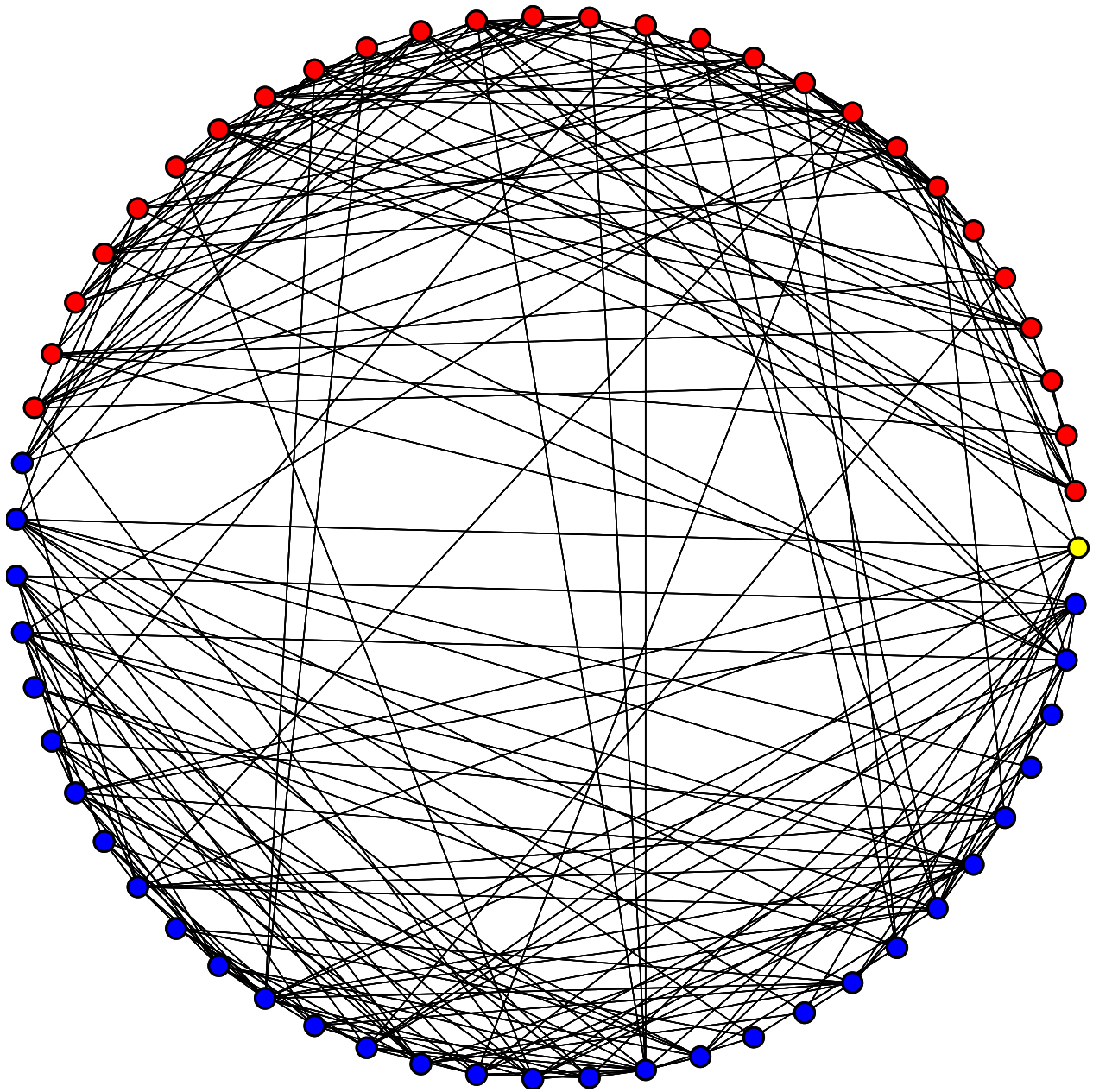
Predict X_s based on $X_{\setminus s} := \{X_s,\ t \neq s\}$.

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

$$\widehat{\theta}[s] \quad := \quad \arg\min_{\theta \in \mathbb{R}^{p-1}} \left\{ \quad -\frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathcal{L}(\theta; X_{i\,\setminus s})}_{\text{local log. likelihood}} \quad + \quad \lambda_n \underbrace{\|\theta\|_1}_{\text{regularization}} \right\}$$

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

US Senate network (2004–2006 voting)



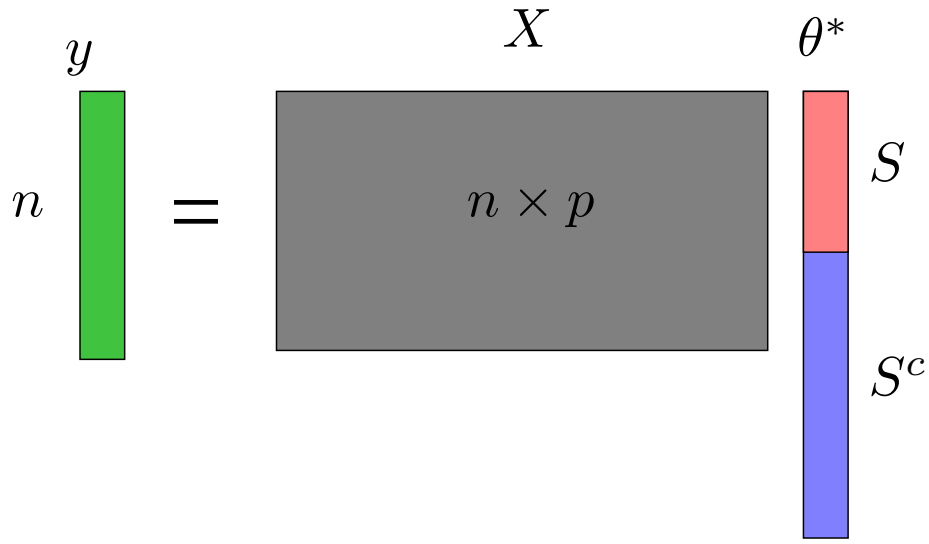
Outline

1. Lecture 1 (Today): Basics of sparse recovery
   - Sparse linear systems: ℓ_0/ℓ_1 equivalence
   - Noisy case: Lasso, ℓ_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 $S \subset \{1, 2, \ldots, p\}$.

ℓ_0-optimization

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

Computationally intractable

NP-hard

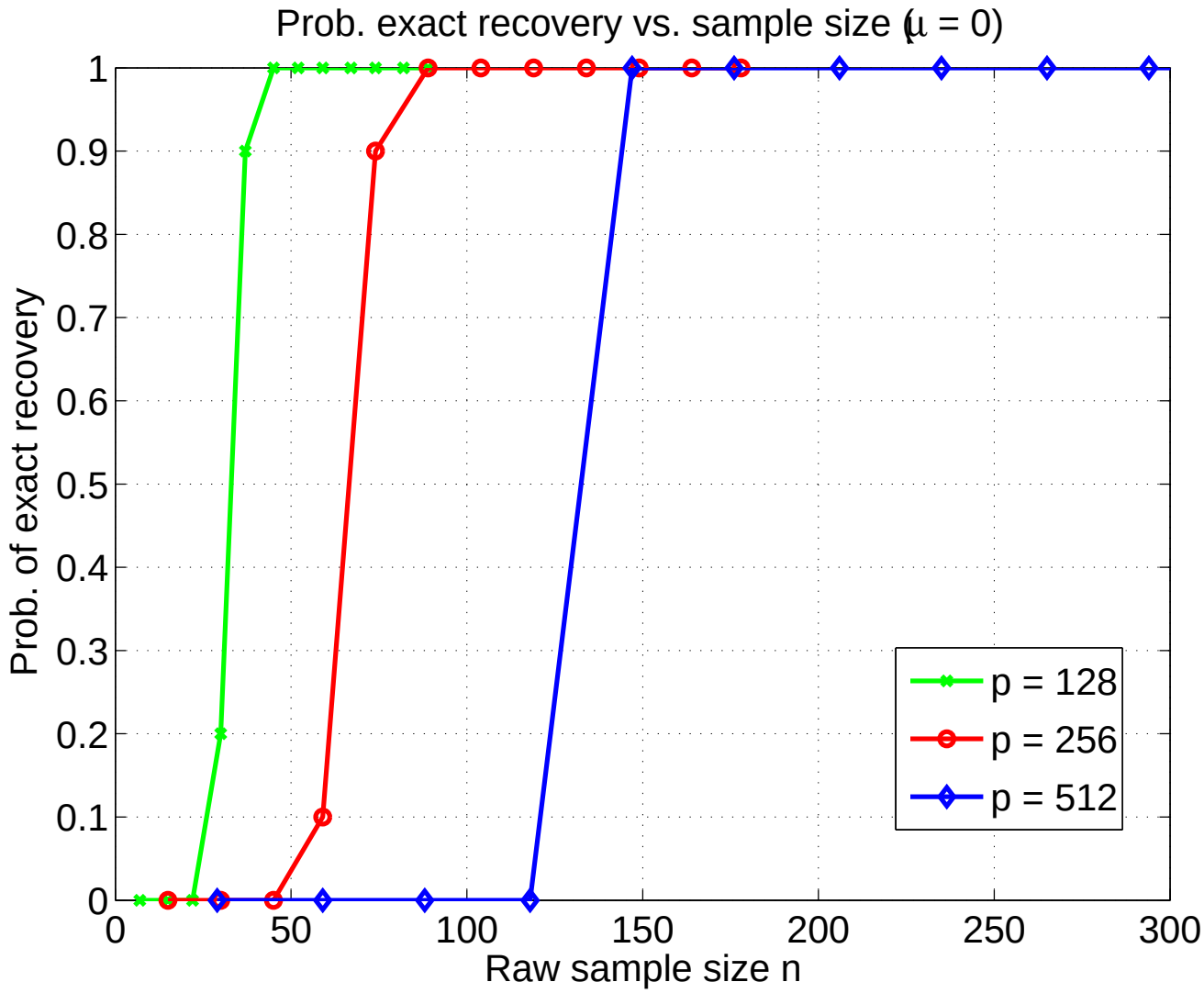
ℓ_1-relaxation

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

Linear program (easy to solve)

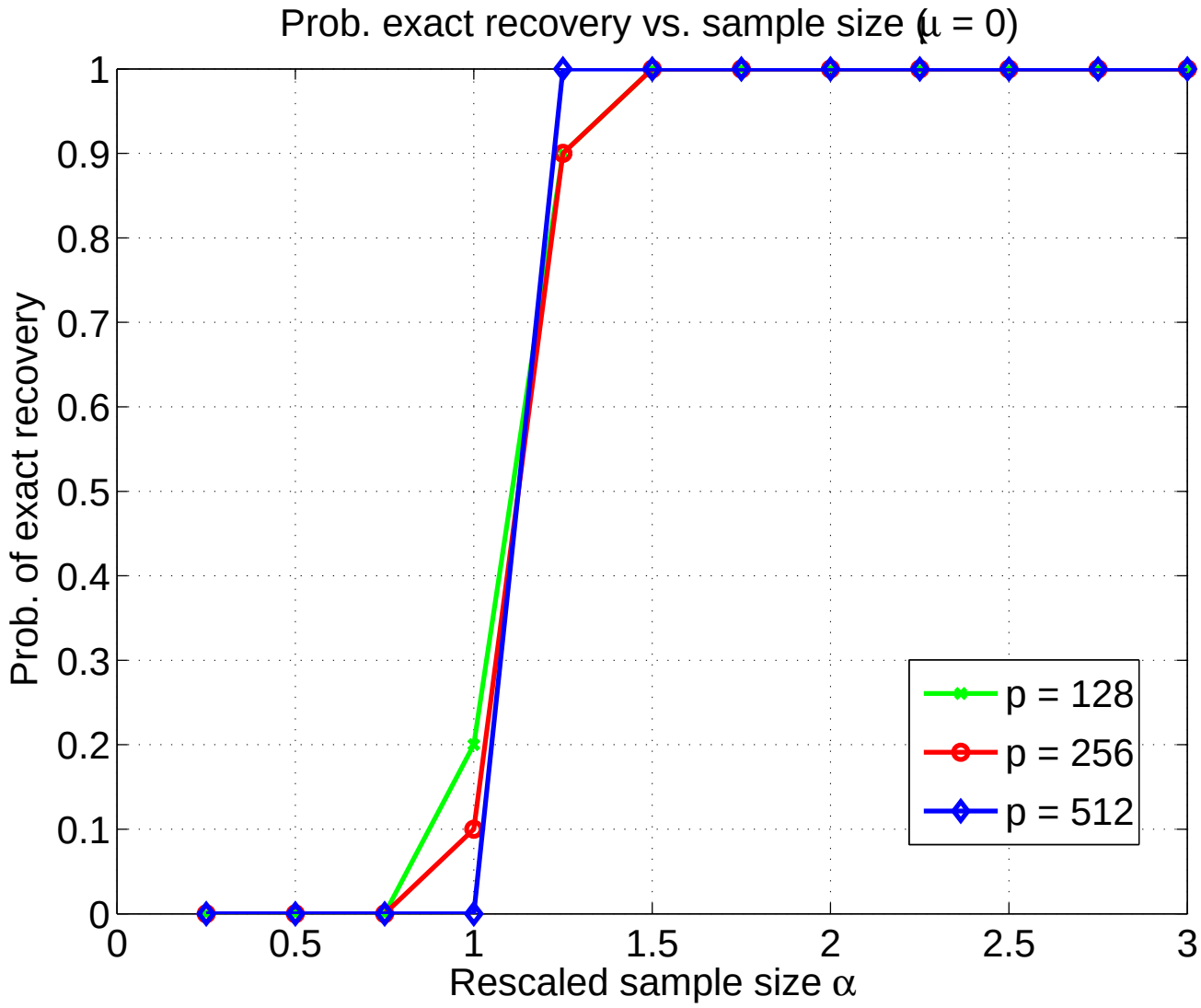
Basis pursuit relaxation

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 $\alpha := \frac{n}{s \log(p/s)}$.

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 $\mathrm{RN}(S)$ for short, if

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

(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 $\mathrm{RN}(S)$ for short, if

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

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

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

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

Proof (sufficiency):

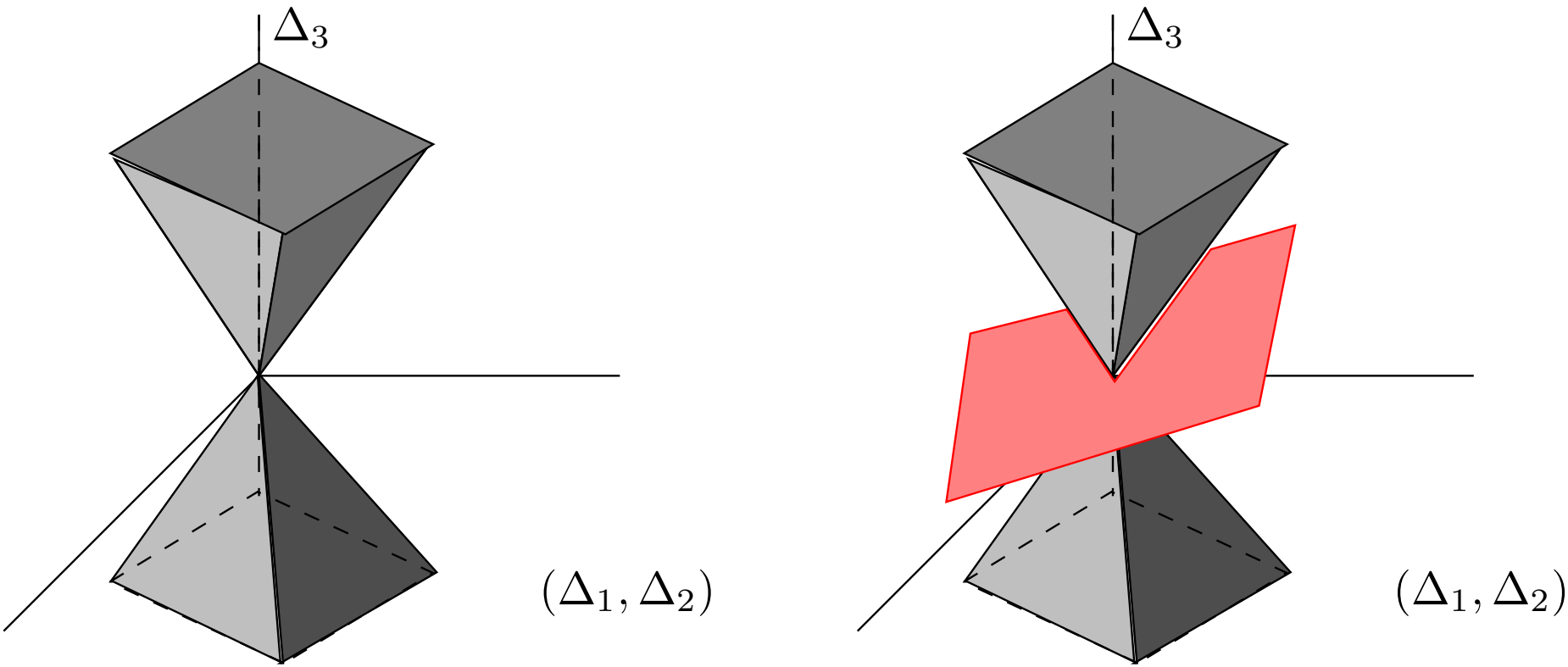
(1) Error vector $\widehat{\Delta} = \theta^* - \widehat{\theta}$ satisfies $X\widehat{\Delta} = 0$, and hence $\widehat{\Delta} \in \mathbb{N}(X)$.

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

$$\begin{aligned}
\text{Optimality of } \widehat{\theta}: \quad & \|\widehat{\theta}\|_1 \leq \|\theta^*\|_1 = \|\theta^*_S\|_1. \\
\text{Sparsity of } \theta^*: \quad & \|\widehat{\theta}\|_1 = \|\theta^* + \widehat{\Delta}\|_1 = \|\theta^*_S + \widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1. \\
\text{Triangle inequality:} \quad & \|\theta^*_S + \widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1 \geq \|\theta^*_S\|_1 - \|\widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1.
\end{aligned}$$

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

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

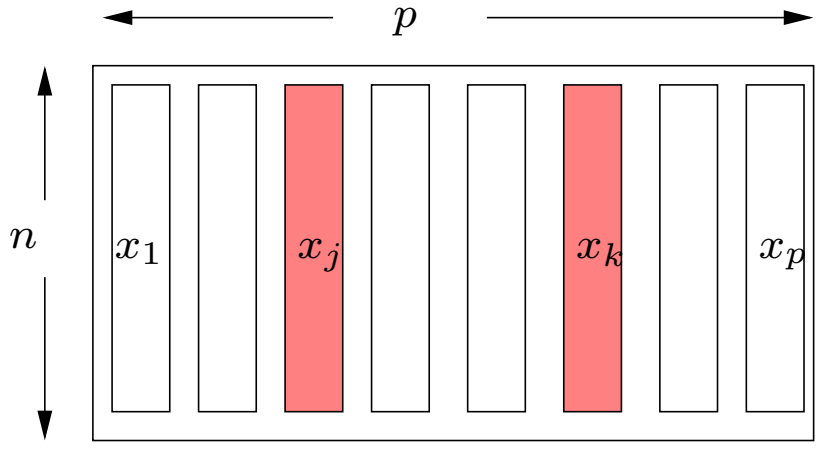
$$\mathbb{C}(S; 1) := \big\{(\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3 \mid |\Delta_1| + |\Delta_2| \leq |\Delta_3|\big\}.$$

Some sufficient conditions

How to verify RN property for a given sparsity s?

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

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

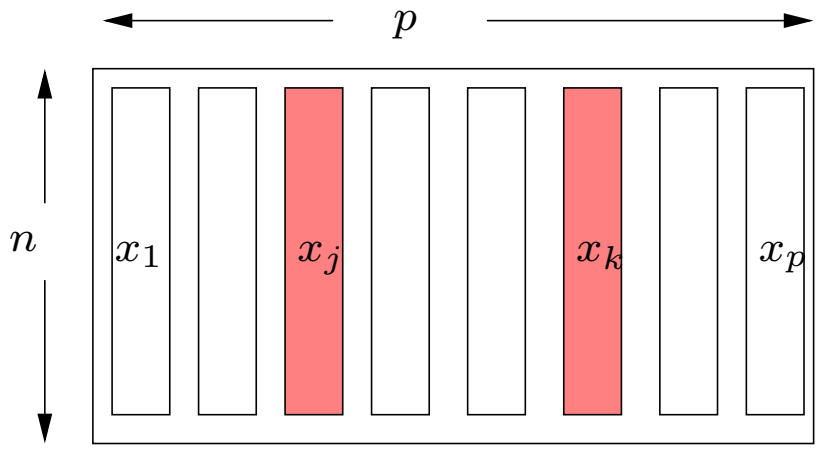


Some sufficient conditions

How to verify RN property for a given sparsity s?

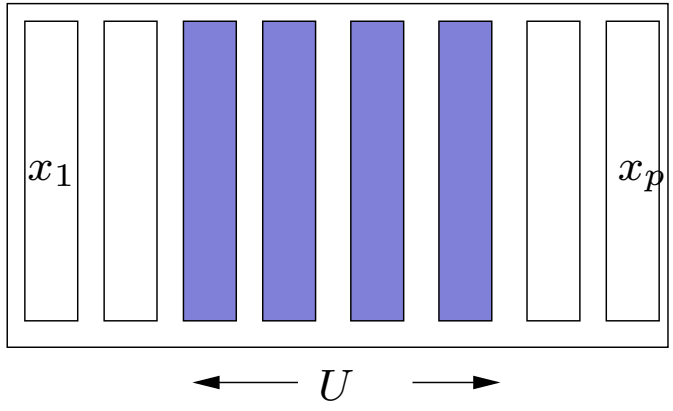
1. Elementwise incoherence condition (Donoho & Xuo, 2001; Feuer & Nem., 2003)

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



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

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

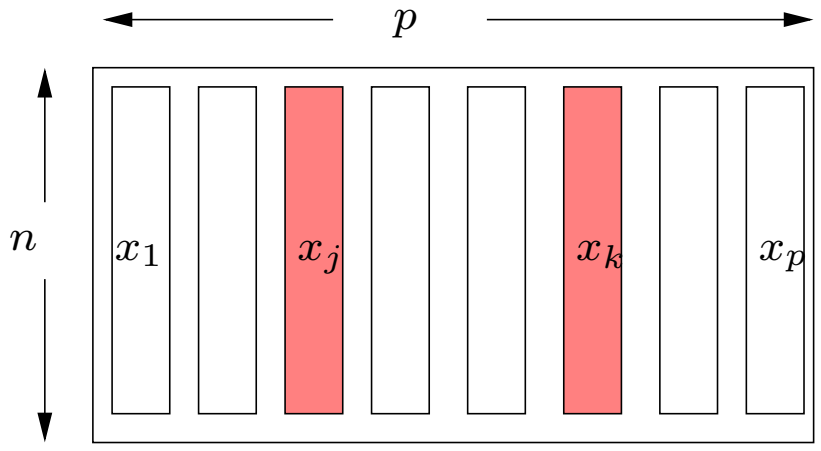


Some sufficient conditions

How to verify RN property for a given sparsity s?

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

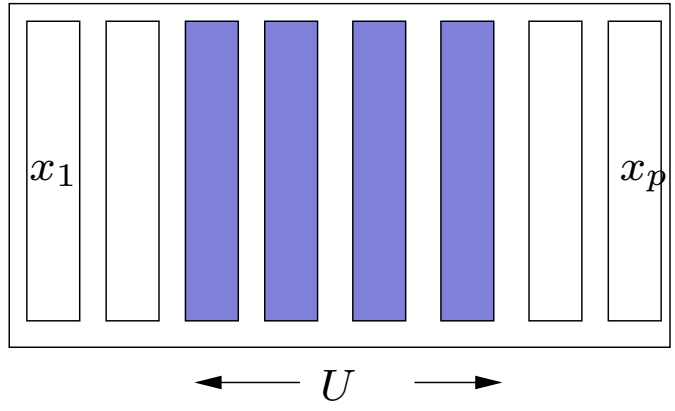
$$\max_{j,k=1,\ldots,p} \left| \left(\frac{X^T X}{n} - I_{p\times p} \right)_{jk} \right| \leq \frac{\delta_1}{s}$$



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)

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

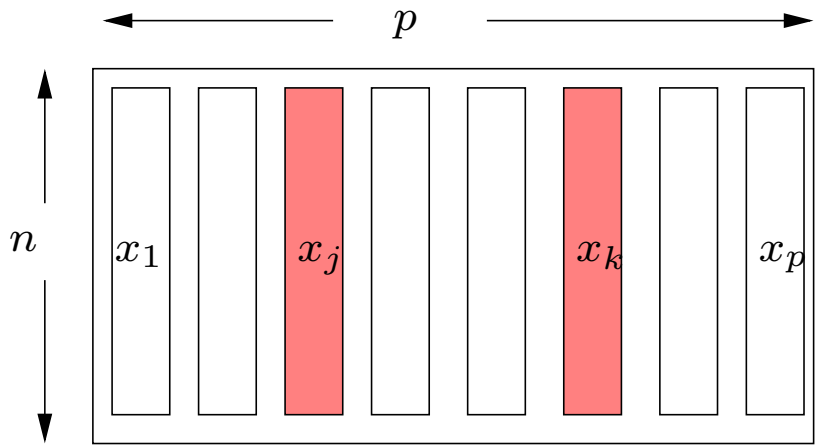


Some sufficient conditions

How to verify RN property for a given sparsity s?

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

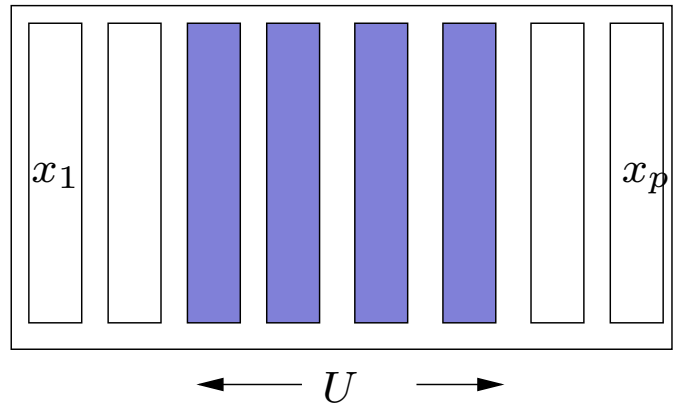
$$\max_{j,k=1,\ldots,p} \left| \left(\frac{X^T X}{n} - I_{p\times p} \right)_{jk} \right| \leq \frac{\delta_1}{s}$$



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)

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



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

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

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

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

Violating matrix incoherence (elementwise/RIP)

Form random design matrix

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

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

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

- 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

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

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

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

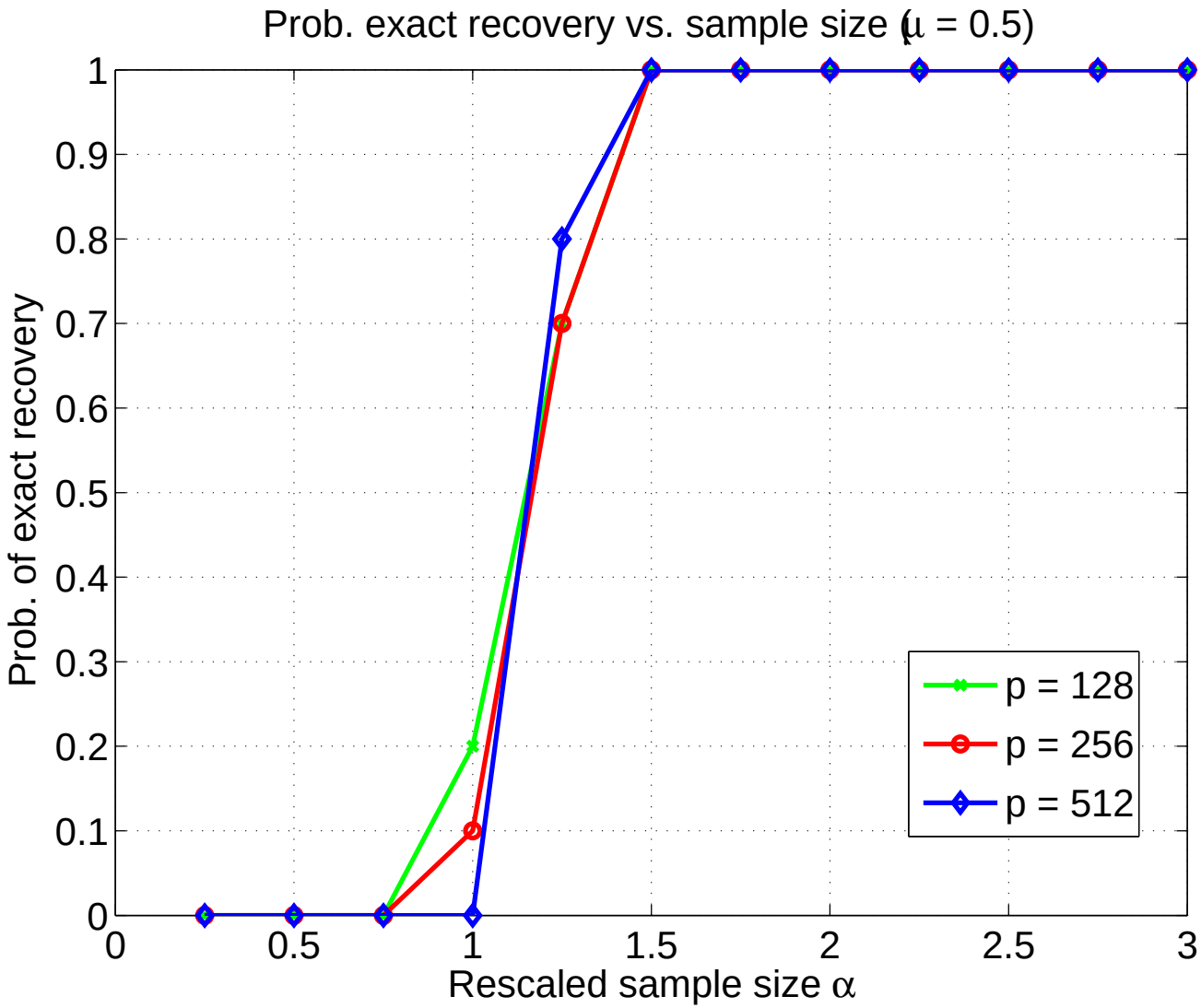
- Elementwise incoherence violated: for any $j \neq k$

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

- RIP constants tend to infinity as $(n, |S|)$ increases:

$$\mathbb{P}\left[\left| \left| \left| \frac{X_S^T X_S}{n} - I_{s \times s} \right| \right|_2 \geq \mu\,(s - 1) - 1 - \epsilon \right] \geq 1 - c_1 \exp(-c_2 n \epsilon^2).$$

Noiseless ℓ_1 recovery for $\mu = 0.5$



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

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* $\kappa(\Sigma) = \max_{j=1,2,\ldots p} \Sigma_{jj}$. *Then for universal constants* c_1, c_2,

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

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* $\kappa(\Sigma) = \max_{j=1,2,\ldots p} \Sigma_{jj}$. *Then for universal constants* c_1, c_2,

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

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

- much less restrictive than incoherence/RIP conditions
- many interesting matrix families are covered
  - Toeplitz dependency
  - constant μ-correlation (previous example)
  - covariance matrix Σ 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

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

- applying previous result:

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

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:

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

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

Easy verification of restricted nullspace

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

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

- applying previous result:

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

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

Definition

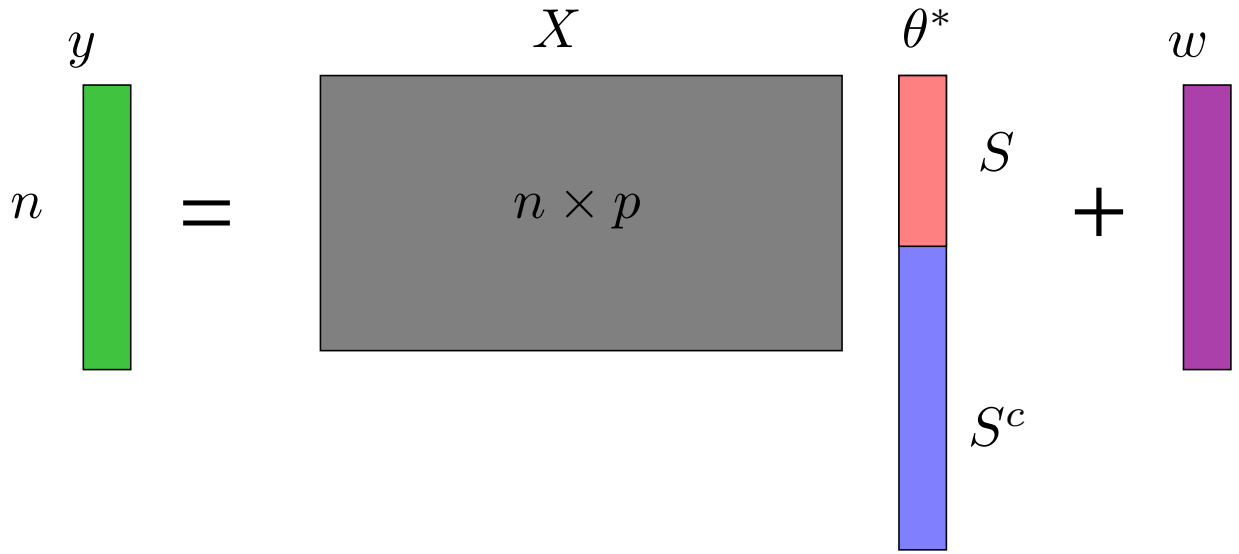
A design matrix $X \in \mathbb{R}^{n\times p}$ satisfies the *restricted eigenvalue* (RE) condition over S (denote RE(S)) with parameters $\alpha \geq 1$ and $\gamma > 0$ if

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

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

Lasso and restricted eigenvalues

Turning to noisy observations...



Estimator: Lasso program

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

Goal: Obtain bounds on $\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2$ that hold with high probability.

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

---

(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 bounds: Four simple steps

Let's analyze constrained version:

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

---

(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 bounds: Four simple steps

Let's analyze constrained version:

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

---

(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

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

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

---

(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

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

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

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

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

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

Look-ahead to Lecture 2: A more general theory

Recap: Thus far.....

- Derived error bounds for basis pursuit and Lasso (ℓ_1-relaxation)
- Seen importance of restricted nullspace and restricted eigenvalues

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 (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)

1. M. J. Wainwright (2009), Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)", *IEEE Trans. Information Theory*, May 2009.

2. S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*. December 2012.

3. G. Raskutti, M. J. Wainwright and B. Yu (2011) Minimax rates for linear regression over ℓ_q-balls. *IEEE Trans. Information Theory*, October 2011.

4. G. Raskutti, M. J. Wainwright and B. Yu (2010). Restricted nullspace and eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*. August 2010.