A primer on high-dimensional statistics: Part I

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Spring School, Westerland, March 2015

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:

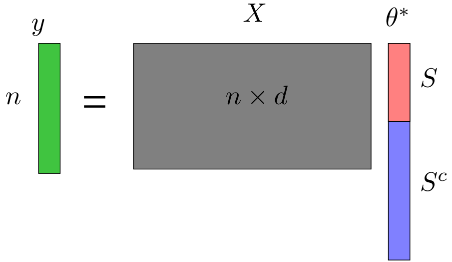
- What embedded low-dimensional structures are present in data?
- How can they can be exploited algorithmically?

Outline

1. Lecture 1—2: Basics of sparse linear models
   - Sparse linear systems: ℓ_0/ℓ_1 equivalence
   - Noisy case: Lasso, ℓ_2-bounds, prediction error and variable selection

2. Lectures 2—3: More general theory

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>

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

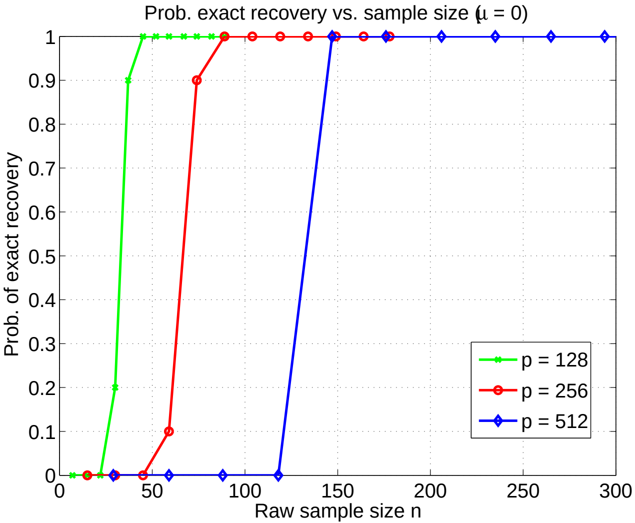
Computationally intractable
NP-hard

<u>ℓ_1-relaxation</u>

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

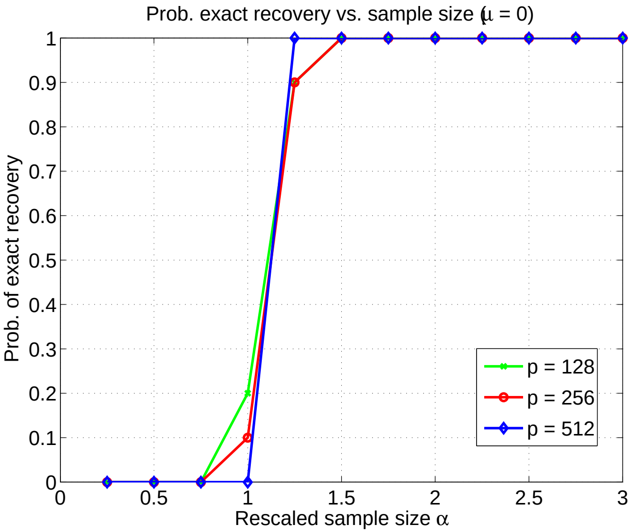
Linear program (easy to solve)
Basis pursuit relaxation

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

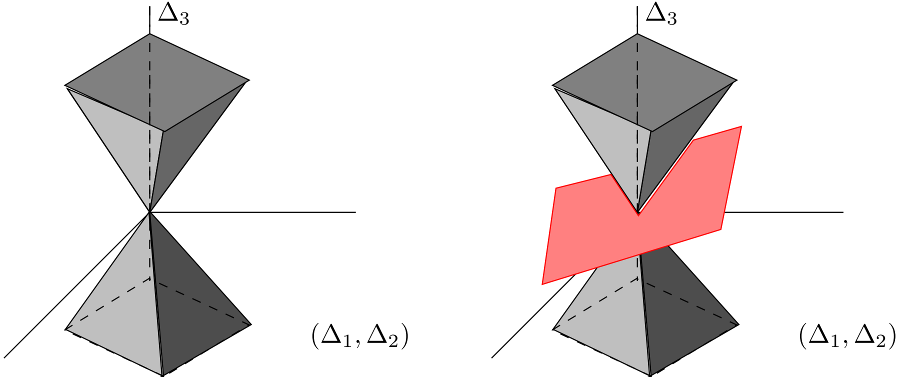
Optimality of $\widehat{\theta}$: $\|\widehat{\theta}\|_1 \leq \|\theta^*\|_1 = \|\theta^*_S\|_1$.

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) $\Longrightarrow \quad \widehat{\Delta} = 0$.

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

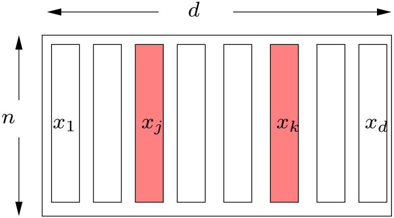
$$\mathbb{C}(S; 1) := \Big\{(\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3 \;\mid\; |\Delta_1| + |\Delta_2| \leq |\Delta_3|\Big\}.$$

Some sufficient conditions

How to verify RN property for a given sparsity s?

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

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

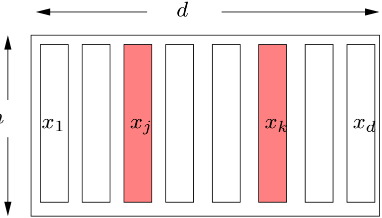


Some sufficient conditions

How to verify RN property for a given sparsity s?

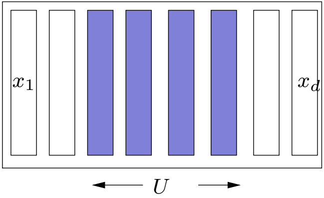
1. Elementwise incoherence condition (Donoho & Xuo, 2001; Feuer & Nem., 2003)

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



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

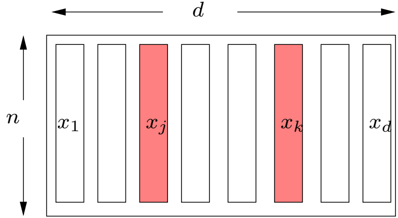
$$\max_{|U| \leq 2s} \left\| \left(\frac{X^T X}{n} - I_{d \times d} \right)_{UU} \right\|_{\text{op}} \leq \delta_{2s}.$$



Some sufficient conditions

How to verify RN property for a given sparsity s?

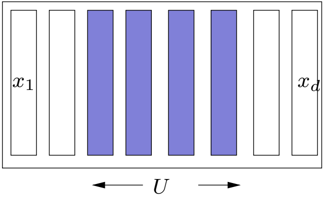
1. Elementwise incoherence condition (Donoho & Xuo, 2001; Feuer & Nem., 2003)



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

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)



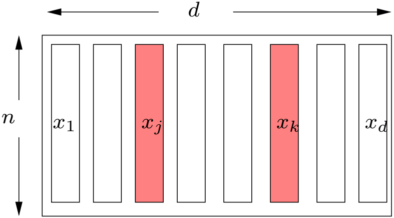
$$\max_{|U|\leq 2s} \left\| \left(\frac{X^T X}{n} - I_{d\times d} \right)_{UU} \right\|_{\mathrm{op}} \leq \delta_{2s}.$$

Some sufficient conditions

How to verify RN property for a given sparsity s?

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

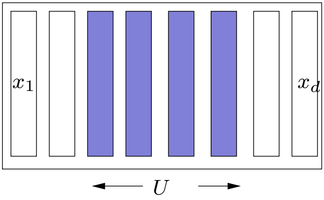
$$\max_{j,k=1,\ldots,d} \left| \left(\frac{X^T X}{n} - I_{d \times d} \right)_{jk} \right| \leq \frac{\delta_1}{s}$$



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

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

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

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

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

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

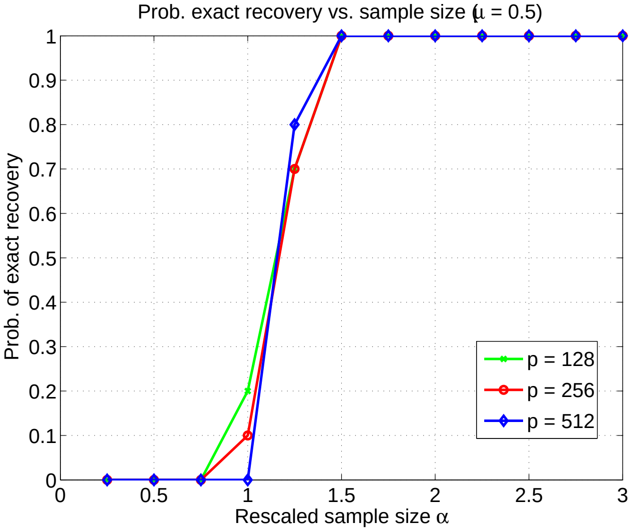
- Elementwise incoherence violated: for any $j \neq k$

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

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

$$\mathbb{P}\Big[\Big\vert\!\Big\vert\!\Big\vert \frac{X_S^T X_S}{n} - I_{s \times s} \Big\vert\!\Big\vert\!\Big\vert_2 \geq \mu \, (s - 1) - 1 - \epsilon \Big] \geq 1 - c_1 \exp(-c_2 n \epsilon^2).$$

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:

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

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:

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

- 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

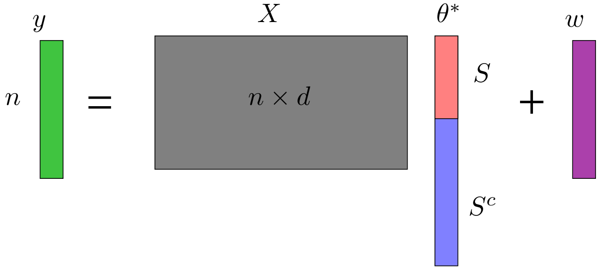
A design matrix $X \in \mathbb{R}^{n \times d}$ satisfies the *restricted eigenvalue* (RE) condition over S (denote $\mathrm{RE}(S)$) with parameters $\alpha \geq 1$ and $\gamma > 0$ if

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

(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 (for slow rate)
   - 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 (for slow rate)
   - 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 (for slow rate)
   - 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 $\mathrm{supp}(\widehat{\theta})$ equal to $\mathrm{supp}(\theta^*)$?
   - 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 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 $\mathrm{RE}(S)$ with parameter $\gamma > 0$.

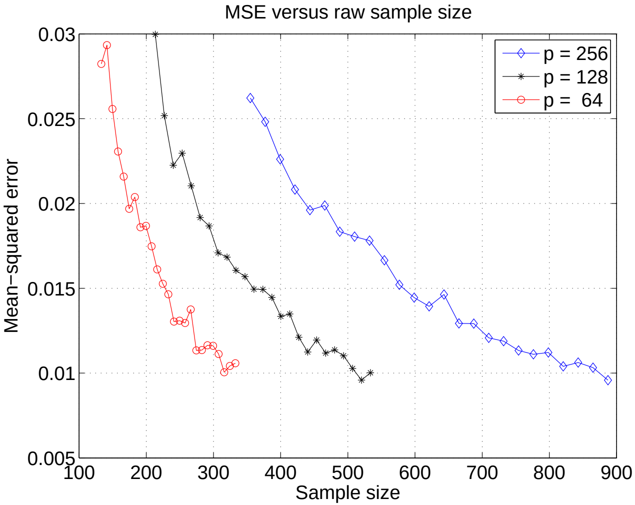
For constrained Lasso with $R = \|\theta^*\|_1$ or regularized Lasso with $\lambda_n = 2\|X^T w/n\|_\infty$, any optimal solution $\widehat{\theta}$ satisfies the bound

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

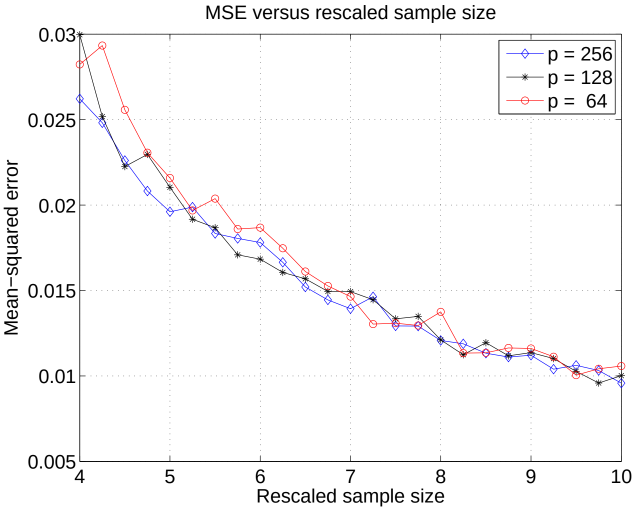
- this is a deterministic result on the set of optimizers
- various corollaries for specific statistical models
  - Compressed sensing: $X_{ij} \sim N(0,1)$ and bounded noise $\|w\|_2 \leq \sigma\sqrt{n}$
  - Deterministic design: X with bounded columns and $w_i \sim N(0, \sigma^2)$

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

Lasso ℓ_2-error: Unrescaled sample size



Lasso ℓ_2-error: Rescaled sample size



Rescaled sample size $\frac{n}{s \log p/s}$.

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

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

- when θ^* is exactly sparse, set $S = \mathrm{supp}(\theta^*)$ 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

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

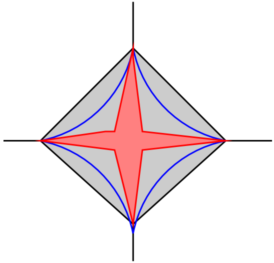
- when θ^* is exactly sparse, set $S = \text{supp}(\theta^*)$ to recover previous result
- more generally, choose S adaptively to trade-off estimation error versus approximation error

(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"

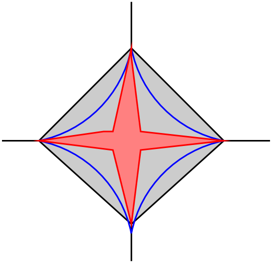
$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q\big\}.$$



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

Bounds on prediction error

Can predict a new response vector $y \in \mathbb{R}^n$ via $\widehat{y} = X\widehat{\theta}$. Associated mean-squared error

$$\frac{1}{n}\mathbb{E}\big[\|y - \widehat{y}\|_2^2\big] = \frac{1}{n}\|X\widehat{\theta} - X\theta^*)\|_2^2 + \sigma^2.$$

Theorem

Consider the constrained Lasso with $R = \|\theta^*\|_1$ or regularized Lasso with $\lambda_n = 4\sigma\sqrt{\frac{\log d}{n}}$ applied to an S-sparse problem with σ-sub-Gaussian noise. Then with high probability:

Slow rate: *If X has normalized columns ($\max_j \|X_j\|_2/\sqrt{n} \leq C$), then any optimal $\widehat{\theta}$ satisfies the bound*

$$\frac{1}{n}\|X\widehat{\theta} - X\theta^*\|_2^2 \leq c\, C\, R\sigma\sqrt{\frac{\log d}{n}}$$

Fast rate: *If X satisfies the γ-RE condition over S, then*

$$\frac{1}{n}\|X\widehat{\theta} - X\theta^*\|_2^2 \leq \frac{c\sigma^2}{\gamma}\,\frac{s\log d}{n}$$

Prediction error: Proof of slow rate

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^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

Prediction error: Proof of slow rate

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^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

Prediction error: Proof of slow rate

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^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

(3) Since both $\widehat{\theta}$ and θ^* are feasible, we have $\|\widehat{\Delta}\|_1 \leq 2R$ by triangle inequality, and hence

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 4R \left\|\frac{X^T w}{n}\right\|_\infty.$$

Prediction error: Proof of fast rate

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^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

Prediction error: Proof of fast rate

(1) By optimality of $\widehat{\theta}$ and feasibility of θ^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

Prediction error: Proof of fast rate

(1) By optimality of $\widehat{\theta}$ and feasibility of θ^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

(3) Since $\widehat{\Delta} \in \mathbb{C}(S)$, we have $\|\widehat{\Delta}\|_1 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2$, and hence

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2\Big\|\frac{X^T w}{n}\Big\|_\infty.$$

Prediction error: Proof of fast rate

(1) By optimality of $\widehat{\theta}$ and feasibility of θ^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

(3) Since $\widehat{\Delta} \in \mathbb{C}(S)$, we have $\|\widehat{\Delta}\|_1 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2$, and hence

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2 \Big\|\frac{X^T w}{n}\Big\|_\infty.$$

(4) Now apply γ-RE condition to RHS

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

Cancel terms and re-arrange.

Why RE conditions for fast rate?

Bothersome issue:

Why should prediction performance depend on an RE-condition?

- it is not fundamental: a method based on ℓ_0-regularization (exponential time) can achieve the fast rate with only column normalization

Why RE conditions for fast rate?

Bothersome issue:

Why should prediction performance depend on an RE-condition?

- it is not fundamental: a method based on ℓ_0-regularization (exponential time) can achieve the fast rate with only column normalization

- some negative evidence: an explicit design matrix and sparse vector ($k = 2$) for which Lasso achieves slow rate Foygel & Srebro (2011)

Why RE conditions for fast rate?

Bothersome issue:

Why should prediction performance depend on an RE-condition?

- it is not fundamental: a method based on ℓ_0-regularization (exponential time) can achieve the fast rate with only column normalization
- some negative evidence: an explicit design matrix and sparse vector ($k = 2$) for which Lasso achieves slow rate Foygel & Srebro (2011)
- ....but adaptive Lasso can achieve the fast rate for this counterexample.

A computationally-constrained minimax rate

Complexity classes:

P:	decision problems solvable in poly. time by a Turing machine
P/poly:	class **P** plus polynomial-length advice string

Asssumptions:

- standard linear regression model $y = X\theta^* + w$ where $w \sim N(0, \sigma^2 I_{n \times n})$
- $\mathbf{NP} \not\subseteq \mathbf{P/poly}$

A computationally-constrained minimax rate

Complexity classes:

P:	decision problems solvable in poly. time by a Turing machine
P/poly:	class **P** plus polynomial-length advice string

Asssumptions:

- standard linear regression model $y = X\theta^* + w$ where $w \sim N(0, \sigma^2 I_{n\times n})$
- **NP** $\not\subseteq$ **P/poly**

Theorem (Zhang, W. & Jordan, COLT 2014)

There is a fixed "bad" design matrix $X \in \mathbb{R}^{n\times d}$ with RE constant $\gamma(X)$ such for any polynomial-time computable $\widehat{\theta}$ returning s-sparse outputs:

$$\sup_{\theta^* \in \mathbb{B}_0(s)} \mathbb{E}\Big[\frac{\|X(\widehat{\theta} - \theta^*)\|_2^2}{n}\Big] \succsim \frac{\sigma^2}{\gamma^2(X)} \frac{s^{1-\delta}\log d}{n}.$$

A primer on high-dimensional statistics: Part II

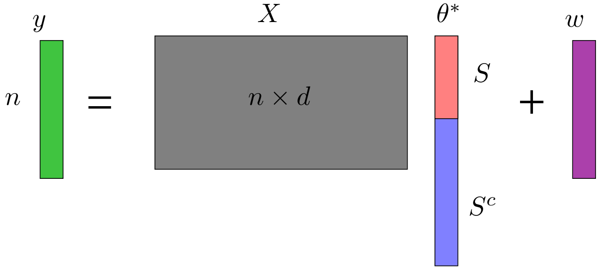
Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Spring School, Westerland, March 2015

Analysis of Lasso estimator

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 (for slow rate)
   - 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 (for slow rate)
   - 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 (for slow rate)
   - 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 $\mathrm{supp}(\widehat{\theta})$ equal to $\mathrm{supp}(\theta^*)$?
   - 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

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

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

(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 $S = \mathrm{supp}(\theta^*)$ 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

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

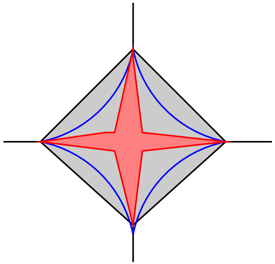
- when θ^* is exactly sparse, set $S = \operatorname{supp}(\theta^*)$ to recover previous result
- more generally, choose S adaptively to trade-off estimation error versus approximation error

(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"

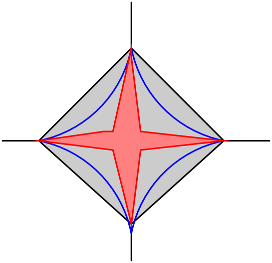
$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \le R_q\big\}.$$



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

Variable selection consistency

Question

When is Lasso solution unique with $\text{support}(\widehat{\theta}) = \text{suppport}(\theta^*)$?

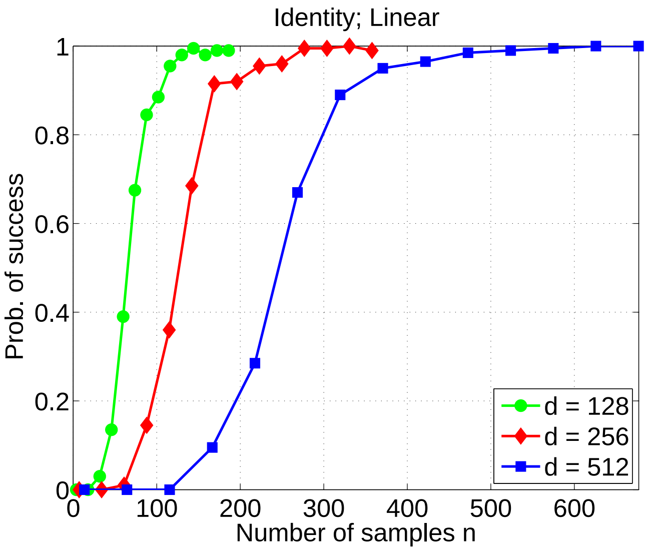
Variable selection consistency

Question

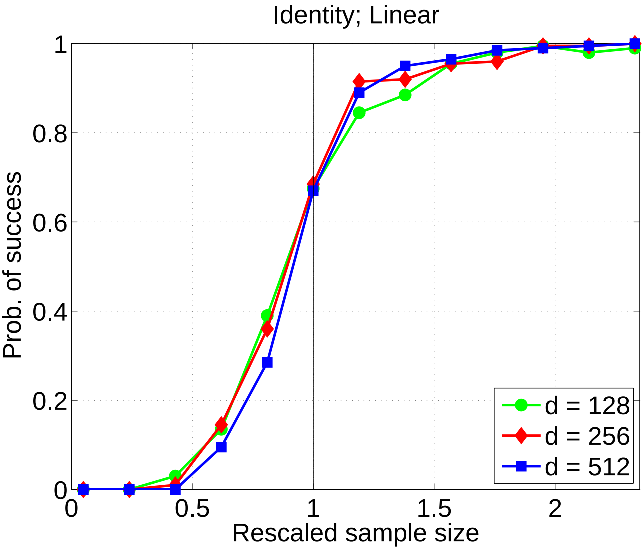
When is Lasso solution unique with $\text{support}(\widehat{\theta}) = \text{suppport}(\theta^*)$?

- Requires a different proof technique, known as a primal-dual witness method.
- A procedure that attempts to construct a pair $(\widehat{\theta}, \widehat{z}) \in \mathbb{R}^d \times \mathbb{R}^d$ that satisfy the KKT conditions for convex optimality
- When procedure succeeds, it certifies the uniqueness and optimality of $\widehat{\theta}$ as a Lasso solution.

Variable selection performance: unrescaled plots



Variable selection performance: rescaled plots



Rescaled sample size: $\frac{n}{s \log(d-s)}$

Primal-dual witness construction

Consider blocks $\begin{bmatrix}\widehat{\theta}_S & \widehat{\theta}_{S^c}\end{bmatrix}$ and $\begin{bmatrix}\widehat{z}_S & \widehat{z}_{S^c}\end{bmatrix}$.

1. Set $\widehat{\theta}_{S^c} = 0$.

Primal-dual witness construction

Consider blocks $\begin{bmatrix}\widehat{\theta}_S & \widehat{\theta}_{S^c}\end{bmatrix}$ and $\begin{bmatrix}\widehat{z}_S & \widehat{z}_{S^c}\end{bmatrix}$.

1. Set $\widehat{\theta}_{S^c} = 0$.
2. Solve oracle sub-problem

$$\widehat{\theta}_S = \arg\min_{\theta_S \in \mathbb{R}^{|S|}} \left\{ \frac{1}{2n} \|y - X_S\theta_S\|_2^2 + \lambda_n \|\theta_S\|_1 \right\},$$

   and choose $\widehat{z} \in \partial \|\theta_S\|_1 \Big|_{\theta_S = \widehat{\theta}_S}$ such that $\frac{1}{n} X_S^T \big(X_s \widehat{\theta}_S - y\big) + \lambda_n \widehat{z}_S = 0$.
   Require $\frac{1}{n}\lambda_{min}(X_S^T X_S) > 0$ in this step.

Primal-dual witness construction

Consider blocks $\begin{bmatrix}\widehat{\theta}_S & \widehat{\theta}_{S^c}\end{bmatrix}$ and $\begin{bmatrix}\widehat{z}_S & \widehat{z}_{S^c}\end{bmatrix}$.

1. Set $\widehat{\theta}_{S^c} = 0$.
2. Solve oracle sub-problem

$$\widehat{\theta}_S = \arg\min_{\theta_S \in \mathbb{R}^{|S|}} \left\{ \frac{1}{2n} \|y - X_S \theta_S\|_2^2 + \lambda_n \|\theta_S\|_1 \right\},$$

   and choose $\widehat{z} \in \partial \|\theta_S\|_1 \Big|_{\theta_S = \widehat{\theta}_S}$ such that $\frac{1}{n} X_S^T \big(X_s \widehat{\theta}_S - y\big) + \lambda_n \widehat{z}_S = 0$.
   Require $\frac{1}{n} \lambda_{min}(X_S^T X_S) > 0$ in this step.
3. Choose $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ to satisfy the zero-subgradient conditions, and such that $\|\widehat{z}_{S^c}\|_\infty < 1$.

Primal-dual witness construction

Consider blocks $\begin{bmatrix}\widehat{\theta}_S & \widehat{\theta}_{S^c}\end{bmatrix}$ and $\begin{bmatrix}\widehat{z}_S & \widehat{z}_{S^c}\end{bmatrix}$.

1. Set $\widehat{\theta}_{S^c} = 0$.
2. Solve oracle sub-problem

$$\widehat{\theta}_S = \arg \min_{\theta_S \in \mathbb{R}^{|S|}} \left\{ \frac{1}{2n} \|y - X_S \theta_S\|_2^2 + \lambda_n \|\theta_S\|_1 \right\},$$

   and choose $\widehat{z} \in \partial \|\theta_S\|_1 \Big|_{\theta_S = \widehat{\theta}_S}$ such that $\frac{1}{n} X_S^T \big(X_s \widehat{\theta}_S - y\big) + \lambda_n \widehat{z}_S = 0$. Require $\frac{1}{n} \lambda_{min}(X_S^T X_S) > 0$ in this step.
3. Choose $\widehat{z}_{S^c} \in \mathbb{R}^{d-s}$ to satisfy the zero-subgradient conditions, and such that $\|\widehat{z}_{S^c}\|_\infty < 1$.

Lemma

If the PDW succeeds, then $\widehat{\theta}$ is the unique optimal solution of the Lasso and satisfies support$(\widehat{\theta}) \subseteq$ support(θ^*).

Proof sketch

1. Form zero-subgradient conditions:

$$\frac{1}{n}\begin{bmatrix} X_S^T X_S & X_{S^c}^T X_S \\ X_{S^c}^T X_S & X_{S^c}^T X_{S^c} \end{bmatrix} \begin{bmatrix} \widehat{\theta}_S - \theta^*_S \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} X_S^T \\ X_{S^c}^T \end{bmatrix} w + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Proof sketch

1. Form zero-subgradient conditions:

$$\frac{1}{n}\begin{bmatrix} X_S^T X_S & X_{S^c}^T X_S \\ X_{S^c}^T X_S & X_{S^c}^T X_{S^c} \end{bmatrix} \begin{bmatrix} \widehat{\theta}_S - \theta_S^* \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} X_S^T \\ X_{S^c}^T \end{bmatrix} w + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

2. Solve for $\widehat{\theta}_S - \theta_S^*$:

$$\underbrace{\widehat{\theta}_S - \theta_S^*}_{U_S} = -(X_S^T X_S)^{-1} X_S^T \mathbf{w} - \lambda_n n (X_S^T X_S)^{-1} z_S.$$

Proof sketch

1. Form zero-subgradient conditions:

$$\frac{1}{n}\begin{bmatrix} X_S^T X_S & X_{S^c}^T X_S \\ X_{S^c}^T X_S & X_{S^c}^T X_{S^c} \end{bmatrix} \begin{bmatrix} \widehat{\theta}_S - \theta_S^* \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} X_S^T \\ X_{S^c}^T \end{bmatrix} w + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

2. Solve for $\widehat{\theta}_S - \theta_S^*$:

$$\underbrace{\widehat{\theta}_S - \theta_S^*}_{U_S} = -(X_S^T X_S)^{-1} X_S^T \mathbf{w} - \lambda_n n (X_S^T X_S)^{-1} z_S.$$

3. Solve for z_{S^c}:

$$z_{S^c} = \underbrace{X_{S^c}^T X_S (X_S^T X_S)^{-1} z_S}_{\mu} + \underbrace{X_{S^c}^T \Big[I - X_S (X_S^T X_S)^{-1} X_S^T \Big] \big(\frac{\mathbf{w}}{\lambda_n\, n}\big)}_{V_{S^c}}.$$

Checking that $\|z_{S^c}\|_\infty < 1$ requires irrepresentability condition

$$\max_{j \in S^c} X_j^T \| X_S (X_S^T X_S)^{-1} \|_1 < \alpha < 1.$$

Bounds on prediction error

Can predict a new response vector $y \in \mathbb{R}^n$ via $\widehat{y} = X\widehat{\theta}$. Associated mean-squared error

$$\frac{1}{n}\mathbb{E}\big[\|y - \widehat{y}\|_2^2\big] = \frac{1}{n}\|X\widehat{\theta} - X\theta^*)\|_2^2 + \sigma^2.$$

Theorem

Consider the constrained Lasso with $R = \|\theta^*\|_1$ or regularized Lasso with $\lambda_n = 4\sigma\sqrt{\frac{\log d}{n}}$ applied to an S-sparse problem with σ-sub-Gaussian noise. Then with high probability:

Slow rate: *If X has normalized columns ($\max_j \|X_j\|_2/\sqrt{n} \le C$), then any optimal $\widehat{\theta}$ satisfies the bound*

$$\frac{1}{n}\|X\widehat{\theta} - X\theta^*\|_2^2 \le c\,C\,R\sigma\sqrt{\frac{\log d}{n}}$$

Fast rate: *If X satisfies the γ-RE condition over S, then*

$$\frac{1}{n}\|X\widehat{\theta} - X\theta^*\|_2^2 \le \frac{c\sigma^2}{\gamma}\,\frac{s\log d}{n}$$

Prediction error: Proof of slow rate

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^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

Prediction error: Proof of slow rate

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^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

Prediction error: Proof of slow rate

Let's analyze constrained version:

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

---

(1) By optimality of $\widehat{\theta}$ and feasibility of θ^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

(3) Since both $\widehat{\theta}$ and θ^* are feasible, we have $\|\widehat{\Delta}\|_1 \le 2R$ by triangle inequality, and hence

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \le 4R\Big\|\frac{X^T w}{n}\Big\|_\infty.$$

Prediction error: Proof of fast rate

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^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

Prediction error: Proof of fast rate

(1) By optimality of $\widehat{\theta}$ and feasibility of θ^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

Prediction error: Proof of fast rate

(1) By optimality of $\widehat{\theta}$ and feasibility of θ^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

(3) Since $\widehat{\Delta} \in \mathbb{C}(S)$, we have $\|\widehat{\Delta}\|_1 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2$, and hence

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2 \Big\|\frac{X^T w}{n}\Big\|_\infty.$$

Prediction error: Proof of fast rate

(1) By optimality of $\widehat{\theta}$ and feasibility of θ^*, we have the basic inequality for $\widehat{\Delta} = \widehat{\theta} - \theta^*$:

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

(2) Hölder's inequality for RHS

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

(3) Since $\widehat{\Delta} \in \mathbb{C}(S)$, we have $\|\widehat{\Delta}\|_1 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2$, and hence

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2 \Big\|\frac{X^T w}{n}\Big\|_\infty.$$

(4) Now apply γ-RE condition to RHS

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

Cancel terms and re-arrange.

Why RE conditions for fast rate?

Bothersome issue:

Why should prediction performance depend on an RE-condition?

- it is not fundamental: a method based on ℓ_0-regularization (exponential time) can achieve the fast rate with only column normalization

Why RE conditions for fast rate?

Bothersome issue:

Why should prediction performance depend on an RE-condition?

- it is not fundamental: a method based on ℓ_0-regularization (exponential time) can achieve the fast rate with only column normalization
- some negative evidence: an explicit design matrix and sparse vector ($k = 2$) for which Lasso achieves slow rate Foygel & Srebro (2011)

Why RE conditions for fast rate?

Bothersome issue:

Why should prediction performance depend on an RE-condition?

- it is not fundamental: a method based on ℓ_0-regularization (exponential time) can achieve the fast rate with only column normalization
- some negative evidence: an explicit design matrix and sparse vector ($k = 2$) for which Lasso achieves slow rate Foygel & Srebro (2011)
- ....but adaptive Lasso can achieve the fast rate for this counterexample.

A computationally-constrained minimax rate

Complexity classes:

P:	decision problems solvable in poly. time by a Turing machine
P/poly:	class **P** plus polynomial-length advice string

Asssumptions:

- standard linear regression model $y = X\theta^* + w$ where $w \sim N(0, \sigma^2 I_{n \times n})$
- $\mathbf{NP} \not\subseteq \mathbf{P/poly}$

A computationally-constrained minimax rate

Complexity classes:

P: decision problems solvable in poly. time by a Turing machine

P/poly: class **P** plus polynomial-length advice string

Asssumptions:

- standard linear regression model $y = X\theta^* + w$ where $w \sim N(0, \sigma^2 I_{n\times n})$
- $\mathbf{NP} \not\subseteq \mathbf{P/poly}$

Theorem (Zhang, W. & Jordan, COLT 2014)

There is a fixed "bad" design matrix $X \in \mathbb{R}^{n\times d}$ with RE constant $\gamma(X)$ such for any polynomial-time computable $\widehat{\theta}$ returning s-sparse outputs:

$$\sup_{\theta^* \in \mathbb{B}_0(s)} \mathbb{E}\Big[\frac{\|X(\widehat{\theta} - \theta^*)\|_2^2}{n}\Big] \succsim \frac{\sigma^2}{\gamma^2(X)} \frac{s^{1-\delta} \log d}{n}.$$

High-level overview

Regularized M-estimators:

Many statistical estimators take the form:

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

High-level overview

Regularized M-estimators:

Many statistical estimators take the 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:

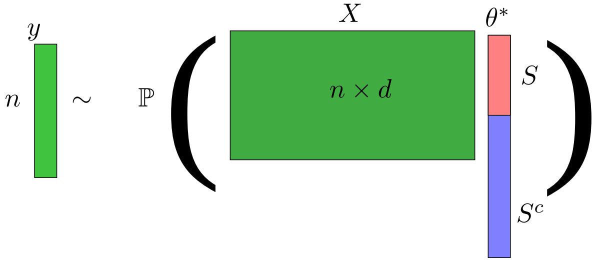
Is there a common set of underlying principles?

Up until now: Sparse regression

Set-up: **Observe** (y_i, x_i) pairs for $i = 1, 2, \ldots, n$, where

$$y_i \sim \mathbb{P}\big(\cdot \mid \langle \theta^*, x_i \rangle\big),$$

where $\theta \in \mathbb{R}^d$ is sparse.

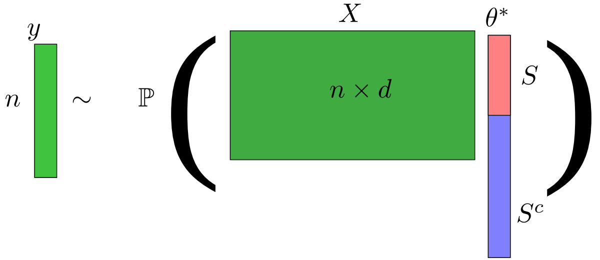


Up until now: Sparse regression

Set-up: **Observe** (y_i, x_i) pairs for $i = 1, 2, \ldots, n$, where

$$y_i \sim \mathbb{P}\big(\cdot \mid \langle \theta^*,\, x_i \rangle\big),$$

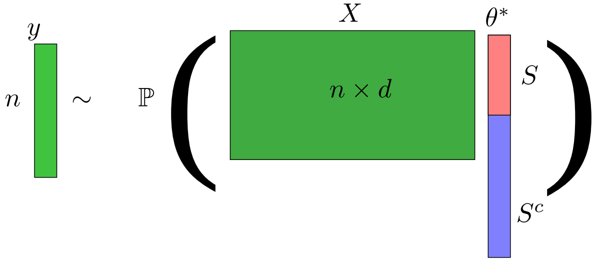
where $\theta \in \mathbb{R}^d$ is sparse.



Estimator: ℓ_1-regularized likelihood

$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ -\frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}\big(y_i \mid \langle x_i,\, \theta \rangle\big) + \lambda \|\theta\|_1 \Big\}.$$

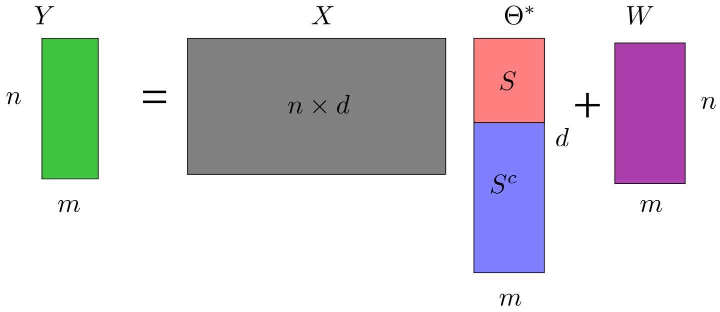
Up until now: Sparse regression



Example: Logistic regression for binary responses $y_i \in \{0, 1\}$:

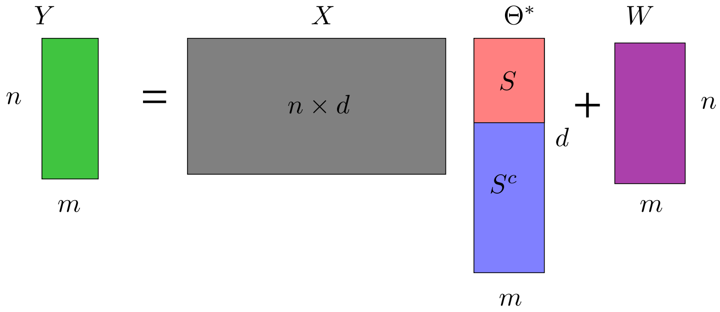
$$\widehat{\theta} \in \arg\min_{\theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \big\{ \log(1 + e^{\langle x_i,\, \theta \rangle}) - y_i \langle x_i,\, \theta \rangle \big\} + \lambda \|\theta\|_1 \Big\}.$$

Example: Block sparsity and group Lasso



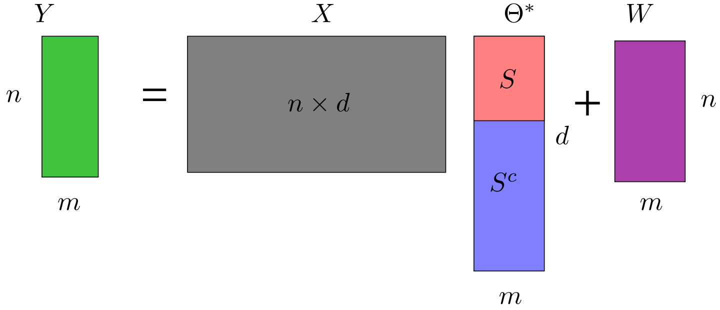
- Matrix Θ^* partitioned into **non-zero rows** S and **zero rows** S^c
- Various applications: multiple-view imaging, gene array prediction, graphical model fitting.

Example: Block sparsity and group Lasso



- Matrix Θ^* partitioned into **non-zero rows** S and **zero rows** S^c
- Various applications: multiple-view imaging, gene array prediction, graphical model fitting.
- Row-wise ℓ_1/ℓ_2-norm $|||\Theta|||_{1,2} = \sum_{j=1}^{d} \|\Theta_j\|_2$

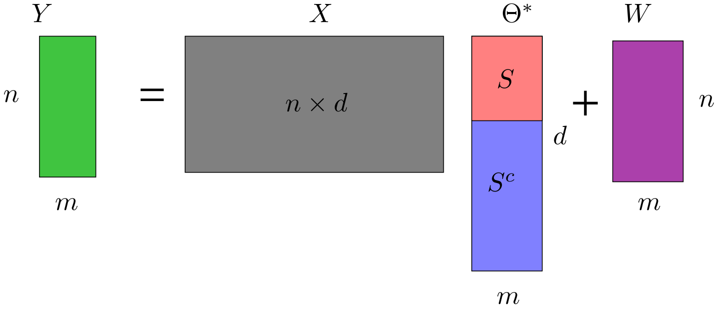
Example: Block sparsity and group Lasso



- Row-wise ℓ_1/ℓ_2-norm $|||\Theta|||_{1,2} = \sum_{j=1}^{d} \|\Theta_j\|_2$
- Weighted r-group Lasso: (Wright et al., 2005; Tropp et al., 2006; Yuan & Lin, 2006)

$$|||\Theta^*|||_{\mathcal{G},r} = \sum_{g \in \mathcal{G}} \omega_g \|\Theta_g\|_r \qquad \text{for some } r \in [2, \infty].$$

Example: Block sparsity and group Lasso

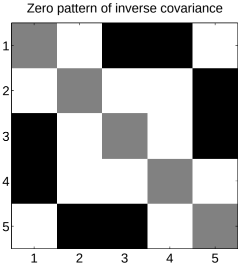


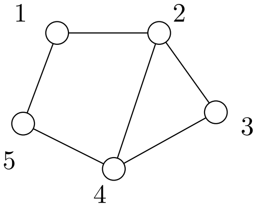
- Row-wise ℓ_1/ℓ_2-norm $|||\Theta|||_{1,2} = \sum_{j=1}^{d} \|\Theta_j\|_2$
- Weighted r-group Lasso: (Wright et al., 2005; Tropp et al., 2006; Yuan & Lin, 2006)

$$|||\Theta^*|||_{\mathcal{G},r} = \sum_{g \in \mathcal{G}} \omega_g \|\Theta_g\|_r \qquad \text{for some } r \in [2, \infty].$$

- Extensions to { hierarchical, graph-based } groups
  (e.g., Zhao et al., 2006; Bach et al., 2009; Baraniuk et al., 2009)

Example: Structured (inverse) covariance matrices





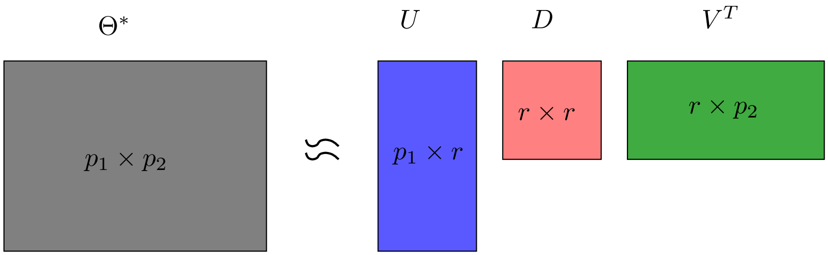
Set-up: Samples from random vector with sparse covariance Σ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{d \times d}$.

Estimator (for inverse covariance)

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

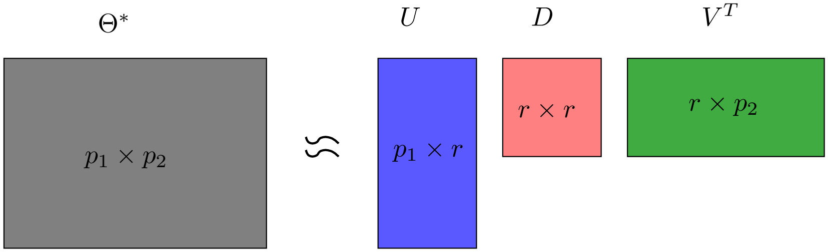
Some past work: Yuan & Lin, 2006; d'Asprémont et al., 2007; Bickel & Levina, 2007; El Karoui, 2007; d'Aspremont et al., 2007; Rothman et al., 2007; Zhou et al., 2007; Friedman et al., 2008; Lam & Fan, 2008; Ravikumar et al., 2008; Zhou, Cai & Huang, 2009; Guo et al., 2010

Example: Low-rank matrix approximation



Set-up: Matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ with rank $r \ll \min\{p_1, p_2\}$.

Example: Low-rank matrix approximation



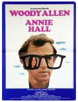
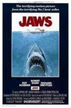
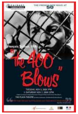
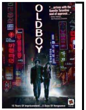
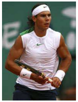
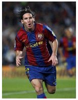
Set-up: Matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ with rank $r \ll \min\{p_1, p_2\}$.

Least-squares matrix regression: Given observations $y_i = \langle\!\langle X_i, \Theta^* \rangle\!\rangle + w_i$, solve:

$$\widehat{\Theta} \in \arg\min_{\Theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\!\langle X_i, \Theta \rangle\!\rangle)^2 + \lambda_n \sum_{j=1}^{\min\{p_1,p_2\}} \gamma_j(\Theta) \Big\}$$

Some past work: Fazel, 2001; Srebro et al., 2004; Recht, Fazel & Parillo, 2007; Bach, 2008; Candes & Recht, 2008; Keshavan et al., 2009; Rohde & Tsybakov, 2010; Recht, 2009; Negahban & W., 2010, Koltchinski et al., 2011

Application: Collaborative filtering

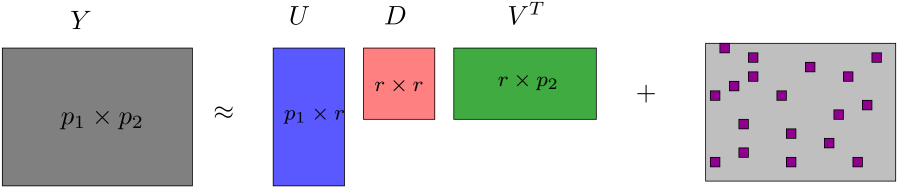
				…	…	
	4	*	3	…	…	*
	3	5	*	…	…	2
	5	4	3	…	…	3
	2	*	*	…	…	1

Universe of p_1 individuals and p_2 films   Observe $n \ll p_2 p_2$ ratings

(e.g., Srebro, Alon & Jaakkola, 2004; Candes & Recht, 2008)

Example: Additive matrix decomposition

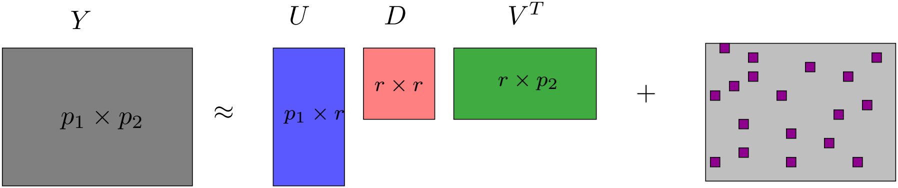
Matrix Y can be (approximately) decomposed into sum:



$$Y = \underbrace{\Theta^*}_{\text{Low-rank component}} + \underbrace{\Gamma^*}_{\text{Sparse component}}$$

Example: Additive matrix decomposition

Matrix Y can be (approximately) decomposed into sum:

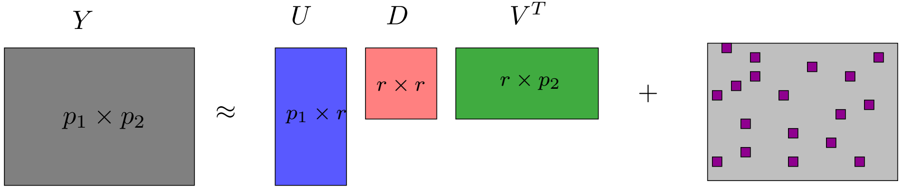


$$Y = \underbrace{\Theta^*}_{\text{Low-rank component}} + \underbrace{\Gamma^*}_{\text{Sparse component}}$$

- Initially proposed by Chandrasekaran, Sanghavi, Parillo & Willsky, 2009
- Various applications:
  - robust collaborative filtering
  - robust PCA
  - graphical model selection with hidden variables

Example: Additive matrix decomposition

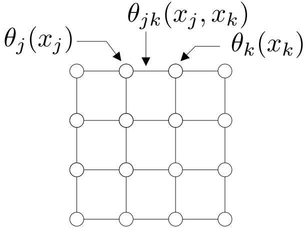
Matrix Y can be (approximately) decomposed into sum:



$$Y = \underbrace{\Theta^*}_{\text{Low-rank component}} + \underbrace{\Gamma^*}_{\text{Sparse component}}$$

- Initially proposed by Chandrasekaran, Sanghavi, Parillo & Willsky, 2009
- Various applications:
  - robust collaborative filtering
  - robust PCA
  - graphical model selection with hidden variables
- subsequent work: Candes et al., 2010; Xu et al., 2010 Hsu et al., 2010; Agarwal et al., 2011

Example: Discrete Markov random fields



Set-up: Samples from discrete MRF(e.g., Ising or Potts model):

$$\mathbb{P}_\theta(x_1, \ldots, x_d) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{j \in V} \theta_j(x_j) + \sum_{(j,k) \in E} \theta_{jk}(x_j, x_k) \Big\}.$$

Estimator: Given empirical marginal distributions $\{\widehat{\mu}_j, \widehat{\mu}_{jk}\}$:

$$\widehat{\Theta} \in \arg\min_{\Theta} \Big\{ \sum_{s \in V} \mathbb{E}_{\widehat{\mu}_j}[\theta_j(x_j)] + \sum_{(j,k)} \mathbb{E}_{\widehat{\mu}_{jk}}[\theta_{jk}(x_j, x_k)] - \log Z(\theta) + \lambda_n \sum_{(j,k)} |||\theta_{jk}|||_F \Big\}$$

Some past work: Spirtes et al., 2001; Abbeel et al., 2005; Csiszar & Telata, 2005; Ravikumar et al, 2007; Schneidman et al., 2007; Santhanam & Wainwright, 2008; Sly et al., 2008; Montanari and Pereira, 2009; Anandkumar et al., 2010

Non-parametric problems: Sparse additive models

- non-parametric regression: severe curse of dimensionality!
- many structured classes of non-parametric models are possible:

Non-parametric problems: Sparse additive models

- non-parametric regression: severe curse of dimensionality!
- many structured classes of non-parametric models are possible:
  - additive models $f^*(x) = \sum_{j=1}^{d} f_j^*(x_j)$ (Stone, 1985)
  - multiple-index models $f^*(x) = g\big(B^* x\big)$

Non-parametric problems: Sparse additive models

- non-parametric regression: severe curse of dimensionality!
- many structured classes of non-parametric models are possible:
  - additive models $f^*(x) = \sum_{j=1}^{d} f_j^*(x_j)$ (Stone, 1985)
  - multiple-index models $f^*(x) = g\big(B^* x\big)$
  - sparse additive models:

$$f^*(x) = \sum_{j \in S}^{d} f_j^*(x_j) \qquad \text{for unknown subset } S$$

(Lin & Zhang, 2003; Meier et al., 2007; Ravikumar et al. 2007; Koltchinski and Yuan, 2008; Raskutti et al., 2010)

Non-parametric problems: Sparse additive models

Sparse additive models:

$$f^*(x) = \sum_{j \in S}^{d} f_j^*(x_j) \qquad \text{for unknown subset } S$$

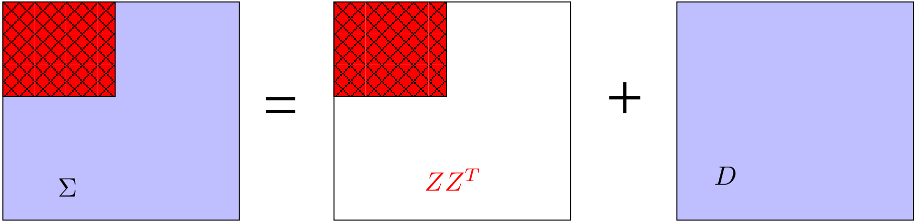
(Lin & Zhang, 2003; Meier et al., 2007; Ravikumar et al. 2007; Koltchinski and Yuan, 2008; Raskutti, W., & Yu, 2010)

Noisy observations $y_i = f^*(x_i) + w_i$ for $i = 1, \ldots, n$.

Estimator:

$$\widehat{f} \in \arg \min_{f = \sum_{j=1}^d f_j} \Big\{ \frac{1}{n} \sum_{i=1}^n \big(y_i - \sum_{j=1}^d f_j(x_{ij}) \big)^2 + \lambda \underbrace{\sum_{j=1}^d \|f_j\|_{\mathcal{H}}}_{\|f_j\|_{1,\mathcal{H}}} + \mu_n \underbrace{\sum_{j=1}^d \|f_j\|_n}_{\|f\|_{1,n}} \Big\}.$$

Example: Sparse principal components analysis



Set-up: Covariance matrix $\Sigma = ZZ^T + D$, where leading eigenspace Z has sparse columns.

Estimator:

$$\widehat{\Theta} \in \arg\min_{\Theta} \Big\{ -\langle\!\langle \Theta, \ \widehat{\Sigma} \rangle\!\rangle + \lambda_n \sum_{(j,k)} |\Theta_{jk}| \Big\}$$

Some past work: Johnstone, 2001; Joliffe et al., 2003; Johnstone & Lu, 2004; Zou et al., 2004; d'Asprémont et al., 2007; Johnstone & Paul, 2008; Amini & Wainwright, 2008; Ma, 2012; Berthet & Rigollet, 2012; Nadler et al., 2012

Motivation and roadmap

- many results on different high-dimensional models
- all based on estimators of the type:

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

Motivation and roadmap

- many results on different high-dimensional models
- all based on estimators of the type:

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

Question:

Is there a common set of underlying principles?

Motivation and roadmap

- many results on different high-dimensional models
- all based on estimators of the type:

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

Question:

Is there a common set of underlying principles?

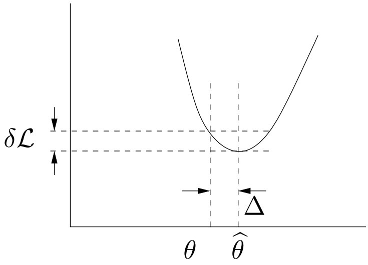
Answer: Yes, two essential ingredients.

(I) Restricted strong convexity of loss function

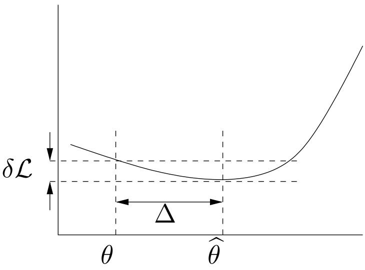
(II) Decomposability of the regularizer

(I) Classical role of curvature in statistics

1. Curvature controls difficulty of estimation:



High curvature: easy to estimate



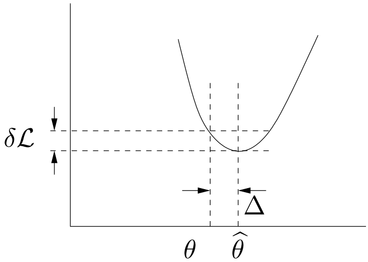
(b) Low curvature: harder

Canonical example:

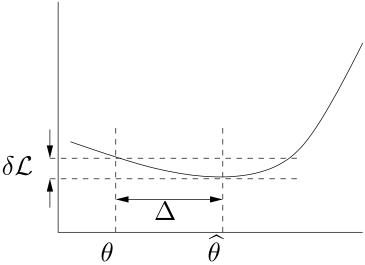
Log likelihood, Fisher information matrix and Cramér-Rao bound.

(I) Classical role of curvature in statistics

1. Curvature controls difficulty of estimation:



High curvature: easy to estimate



(b) Low curvature: harder

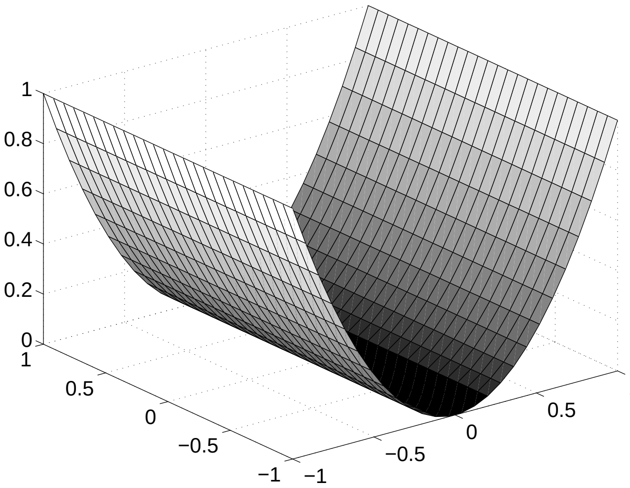
Canonical example:

Log likelihood, Fisher information matrix and Cramér-Rao bound.

2. Formalized by lower bound on Taylor series error $\mathcal{E}_n(\Delta)$

$$\underbrace{\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \, \Delta \rangle}_{\mathcal{E}_n(\Delta)} \geq \gamma^2 \|\Delta\|_\star^2 \quad \text{for all } \Delta \text{ around } \theta^*.$$

High dimensions: no strong convexity!



When $d > n$, the Hessian $\nabla^2 \mathcal{L}(\theta; Z_1^n)$ has nullspace of dimension $d - n$.

Restricted strong convexity

Definition

Loss function $\mathcal{L}_n$ satisfies restricted strong convexity (RSC) with respect to regularizer $\mathcal{R}$ if

$$\underbrace{\mathcal{L}_n(\theta^* + \Delta) - \Big\{\mathcal{L}_n(\theta^*) + \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle\Big\}}_{\text{Taylor error } \mathcal{E}_n(\Delta)} \;\geq\; \underbrace{\gamma_\ell \|\Delta\|_2^2}_{\text{Lower curvature}} \;-\; \underbrace{\tau_\ell^2\, \mathcal{R}^2(\Delta)}_{\text{Tolerance}}$$

for all Δ in a suitable neighborhood of θ^*.

Restricted strong convexity

Definition

Loss function $\mathcal{L}_n$ satisfies restricted strong convexity (RSC) with respect to regularizer $\mathcal{R}$ if

$$\underbrace{\mathcal{L}_n(\theta^* + \Delta) - \Big\{\mathcal{L}_n(\theta^*) + \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle\Big\}}_{\text{Taylor error } \mathcal{E}_n(\Delta)} \geq \underbrace{\gamma_\ell \|\Delta\|_2^2}_{\text{Lower curvature}} - \underbrace{\tau_\ell^2\, \mathcal{R}^2(\Delta)}_{\text{Tolerance}}$$

for all Δ in a suitable neighborhood of θ^*.

- ordinary strong convexity:
  - special case with tolerance $\tau_\ell = 0$
  - does not hold for most loss functions when $d > n$

Restricted strong convexity

Definition

Loss function $\mathcal{L}_n$ satisfies restricted strong convexity (RSC) with respect to regularizer $\mathcal{R}$ if

$$\underbrace{\mathcal{L}_n(\theta^* + \Delta) - \Big\{\mathcal{L}_n(\theta^*) + \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle\Big\}}_{\text{Taylor error } \mathcal{E}_n(\Delta)} \;\geq\; \underbrace{\gamma_\ell \|\Delta\|_2^2}_{\text{Lower curvature}} \;-\; \underbrace{\tau_\ell^2\, \mathcal{R}^2(\Delta)}_{\text{Tolerance}}$$

for all Δ in a suitable neighborhood of θ^*.

- ordinary strong convexity:
  - special case with tolerance $\tau_\ell = 0$
  - does not hold for most loss functions when $d > n$
- RSC enforces a lower bound on curvature, but only when $\mathcal{R}^2(\Delta) \ll \|\Delta\|_2^2$

Restricted strong convexity

Definition

Loss function $\mathcal{L}_n$ satisfies restricted strong convexity (RSC) with respect to regularizer $\mathcal{R}$ if

$$\underbrace{\mathcal{L}_n(\theta^* + \Delta) - \Big\{\mathcal{L}_n(\theta^*) + \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle\Big\}}_{\text{Taylor error } \mathcal{E}_n(\Delta)} \geq \underbrace{\gamma_\ell \|\Delta\|_2^2}_{\text{Lower curvature}} - \underbrace{\tau_\ell^2\, \mathcal{R}^2(\Delta)}_{\text{Tolerance}}$$

for all Δ in a suitable neighborhood of θ^*.

- ordinary strong convexity:
  - special case with tolerance $\tau_\ell = 0$
  - does not hold for most loss functions when $d > n$
- RSC enforces a lower bound on curvature, but only when $\mathcal{R}^2(\Delta) \ll \|\Delta\|_2^2$
- a function satisfying RSC can actually be non-convex

Example: RSC $\equiv$ RE for least-squares

- for least-squares loss $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$:

$$\mathcal{E}_n(\Delta) = \mathcal{L}_n(\theta^* + \Delta) - \Big\{\mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta \rangle\Big\} \;=\; \frac{1}{2n}\|X\Delta\|_2^2.$$

Example: RSC $\equiv$ RE for least-squares

- for least-squares loss $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$:

$$\mathcal{E}_n(\Delta) = \mathcal{L}_n(\theta^* + \Delta) - \Big\{\mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle\Big\} \;=\; \frac{1}{2n}\|X\Delta\|_2^2.$$

- Restricted eigenvalue (RE) condition (van de Geer, 2007; Bickel et al., 2009):

$$\frac{\|X\Delta\|_2^2}{2n} \geq \gamma\, \|\Delta\|_2^2 \quad \text{for all } \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1.$$

Example: RSC $\equiv$ RE for least-squares

- for least-squares loss $\mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$:

$$\mathcal{E}_n(\Delta) = \mathcal{L}_n(\theta^* + \Delta) - \Big\{\mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle\Big\} \;=\; \frac{1}{2n}\|X\Delta\|_2^2.$$

- Restricted eigenvalue (RE) condition (van de Geer, 2007; Bickel et al., 2009):

$$\frac{\|X\Delta\|_2^2}{2n} \geq \gamma\,\|\Delta\|_2^2 \quad \text{for all } \Delta \in \mathbb{R}^d \text{ with } \|\Delta\|_1 \leq 2\sqrt{s}\|\Delta\|_2.$$

Example: Generalized linear models

A broad class of models for relationship between response $y \in \mathcal{X}$ and predictors $x \in \mathbb{R}^d$.

Example: Generalized linear models

A broad class of models for relationship between response $y \in \mathcal{X}$ and predictors $x \in \mathbb{R}^d$.

Based on families of conditional distributions:

$$\mathbb{P}_\theta(y \mid x, \theta^*) \propto \exp\Big\{\frac{y\,\langle x,\,\theta^*\rangle - \Phi(\langle x,\,\theta^*\rangle)}{c(\sigma)}\Big\}.$$

Example: Generalized linear models

A broad class of models for relationship between response $y \in \mathcal{X}$ and predictors $x \in \mathbb{R}^d$.

Based on families of conditional distributions:

$$\mathbb{P}_\theta(y \mid x, \theta^*) \propto \exp\Big\{\frac{y\,\langle x,\,\theta^*\rangle - \Phi(\langle x,\,\theta^*\rangle)}{c(\sigma)}\Big\}.$$

Examples:

- Linear Gaussian model: $\Phi(t) = t^2/2$ and $c(\sigma) = \sigma^2$.
- Binary response data $y \in \{0, 1\}$, Bernouilli model: $\Phi(t) = \log(1 + e^t)$.
- Multinomial responses (e.g., ratings)
- Poisson models (count-valued data): $\Phi(t) = e^t$.

GLM-based restricted strong convexity

- let $\mathcal{R}$ be norm-based regularizer dominating the ℓ_2-norm (e.g., ℓ_1, group-sparse, nuclear etc.)
- let $\mathcal{R}^*$ be the associated dual norm
- covariate-Rademacher complexity of norm ball

$$\sup_{\mathcal{R}(u)\leq 1} \langle u, \, \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i x_i \rangle \;=\; \mathcal{R}^*\big(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i x_i\big)$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d sign variables

GLM-based restricted strong convexity

- let $\mathcal{R}$ be norm-based regularizer dominating the ℓ_2-norm (e.g., ℓ_1, group-sparse, nuclear etc.)
- let $\mathcal{R}^*$ be the associated dual norm
- covariate-Rademacher complexity of norm ball

$$\sup_{\mathcal{R}(u) \leq 1} \langle u, \, \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_i \rangle \; = \; \mathcal{R}^*\big(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_i\big)$$

where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d sign variables

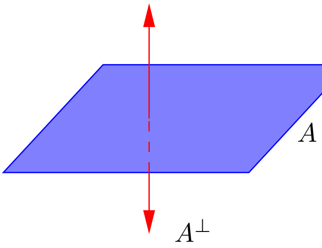
Theorem (Negahban et al., 2012; W. 2014)

Let the covariates $\{x_i\}_{i=1}^n$ be sampled i.i.d. Then

$$\underbrace{\mathcal{E}_n(\Delta)}_{\text{Emp. Taylor error}} \quad \geq \quad \underbrace{\bar{\mathcal{E}}(\Delta)}_{\text{Pop. Taylor error}} \quad - c_1 \big\{t \, \mathcal{R}(\Delta)\big\}^2 \qquad \text{for all } \|\Delta\|_2 \leq 1$$

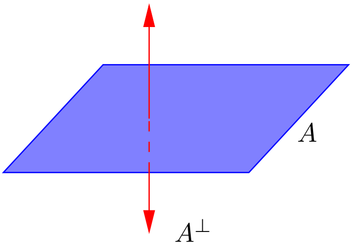
with probability at least $1 - \mathbb{P}[\mathcal{R}^*(\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i) \geq t]$.

(II) Decomposable regularizers



Subspace A:	Approximation to model parameters
Complementary subspace $A^{\perp}$:	Undesirable deviations.

(II) Decomposable regularizers

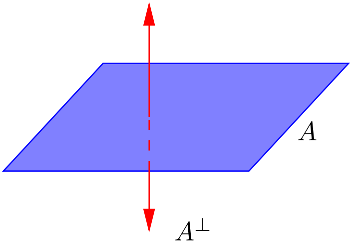


Subspace A:	Approximation to model parameters
Complementary subspace $A^{\perp}$:	Undesirable deviations.

Regularizer $\mathcal{R}$ decomposes across $(A, A^{\perp})$ if

$$\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta) \qquad \text{for all } \alpha \in A, \text{ and } \beta \in A^{\perp}.$$

(II) Decomposable regularizers

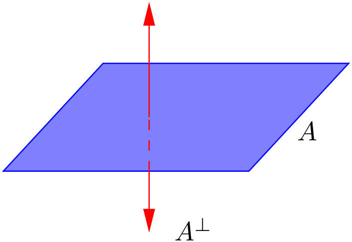


Regularizer $\mathcal{R}$ decomposes across $(A, A^{\perp})$ if

$$\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta) \qquad \text{for all } \alpha \in A, \text{ and } \beta \in A^{\perp}.$$

Includes:
- (weighted) ℓ_1-norms
- group-sparse norms
- nuclear norm
- sums of decomposable norms

(II) Decomposable regularizers



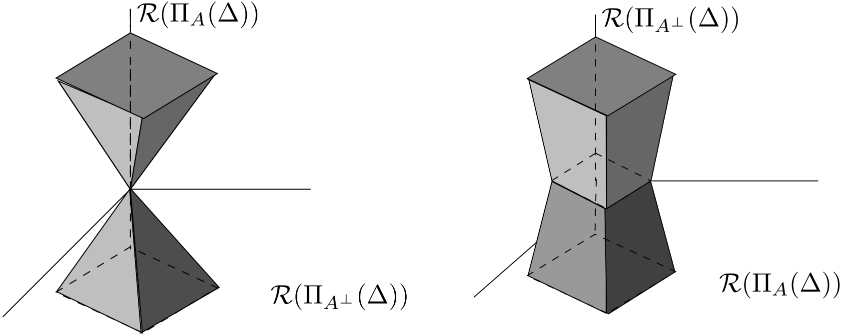
Regularizer $\mathcal{R}$ decomposes across $(A, A^{\perp})$ if

$$\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta) \qquad \text{for all } \alpha \in A, \text{ and } \beta \in A^{\perp}.$$

Related definitions:

Geometric decomposability: Candes & Recht, 2012; Chandrasekaran et al., 20
Weak decomposability: van de Geer, 2012

Significance of decomposability



(a) $\mathbb{C}$ for exact model (cone) (b) $\mathbb{C}$ for approximate model (star-shaped)

Lemma

Suppose that $\mathcal{L}$ is convex, and $\mathcal{R}$ is decomposable w.r.t. A. Then as long as $\lambda_n \geq 2\mathcal{R}^*\Big(\nabla\mathcal{L}(\theta^*; Z_1^n)\Big)$, the error vector $\widehat{\Delta} = \widehat{\theta}_{\lambda_n} - \theta^*$ belongs to

$$\mathbb{C}(A, B; \theta^*) := \big\{\Delta \in \Omega \mid \mathcal{R}(\Pi_{A^\perp}(\Delta)) \leq 3\mathcal{R}(\Pi_B(\Delta)) + 4\mathcal{R}(\Pi_{A^\perp}(\theta^*))\big\}.$$

Example: Sparse vectors and ℓ_1-regularization

- for each subset $S \subset \{1, \ldots, d\}$, define subspace pairs

$$\begin{aligned} A(S) &:= \big\{\theta \in \mathbb{R}^d \mid \theta_{S^c} = 0\big\}, \\ B^\perp(S) &:= \big\{\theta \in \mathbb{R}^d \mid \theta_S = 0\big\} = A^\perp(S). \end{aligned}$$

- decomposability of ℓ_1-norm:

$$\big\|\theta_S + \theta_{S^c}\big\|_1 = \|\theta_S\|_1 + \|\theta_{S^c}\|_1 \quad \text{for all } \theta_S \in A(S) \text{ and } \theta_{S^c} \in B^\perp(S).$$

- natural extension to group Lasso:
  - collection of groups $\mathcal{G}_j$ that partition $\{1, \ldots, d\}$
  - group norm

$$\|\theta\|_{\mathcal{G},\alpha} = \sum_j \|\theta_{\mathcal{G}_j}\|_\alpha \qquad \text{for some } \alpha \in [1, \infty].$$

Example: Low-rank matrices and nuclear norm

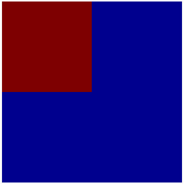
- for each pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{p_1}$ and $V \subseteq \mathbb{R}^{p_2}$:

$$\begin{aligned} A(U,V) &:= \big\{\Theta \in \mathbb{R}^{p_1 \times p_2} \mid \mathrm{row}(\Theta) \subseteq V,\ \mathrm{col}(\Theta) \subseteq U\big\} \\ B^{\perp}(U,V) &:= \big\{\Gamma \in \mathbb{R}^{p_1 \times p_2} \mid \mathrm{row}(\Gamma) \subseteq V^{\perp},\ \mathrm{col}(\Gamma) \subseteq U^{\perp}\big\}. \end{aligned}$$

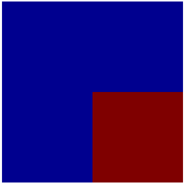
Example: Low-rank matrices and nuclear norm

- for each pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{p_1}$ and $V \subseteq \mathbb{R}^{p_2}$:

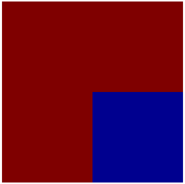
$$\begin{aligned} A(U,V) &:= \big\{\Theta \in \mathbb{R}^{p_1 \times p_2} \mid \mathrm{row}(\Theta) \subseteq V,\ \mathrm{col}(\Theta) \subseteq U\big\} \\ B^{\perp}(U,V) &:= \big\{\Gamma \in \mathbb{R}^{p_1 \times p_2} \mid \mathrm{row}(\Gamma) \subseteq V^{\perp},\ \mathrm{col}(\Gamma) \subseteq U^{\perp}\big\}. \end{aligned}$$



(a) $\Theta \in A$



(b) $\Gamma \in B^{\perp}$

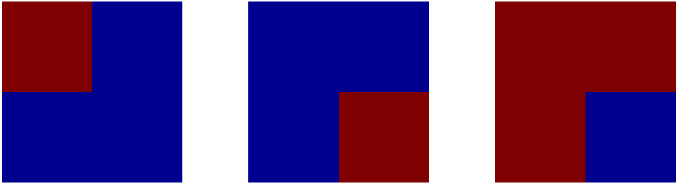


(c) $\Sigma \in B$

Example: Low-rank matrices and nuclear norm

- for each pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{p_1}$ and $V \subseteq \mathbb{R}^{p_2}$:

$$\begin{aligned} A(U,V) &:= \big\{\Theta \in \mathbb{R}^{p_1 \times p_2} \mid \operatorname{row}(\Theta) \subseteq V, \; \operatorname{col}(\Theta) \subseteq U\big\} \\ B^{\perp}(U,V) &:= \big\{\Gamma \in \mathbb{R}^{p_1 \times p_2} \mid \operatorname{row}(\Gamma) \subseteq V^{\perp}, \; \operatorname{col}(\Gamma) \subseteq U^{\perp}\big\}. \end{aligned}$$



(a) $\Theta \in A$ (b) $\Gamma \in B^{\perp}$ (c) $\Sigma \in B$

- by construction, $\Theta^T \Gamma = 0$ for all $\Theta \in A(U,V)$ and $\Gamma \in B^{\perp}(U,V)$
- decomposability of nuclear norm $|||\Theta|||_1 = \sum_{j=1}^{\min\{p_1,p_2\}} \sigma_j(\Theta)$:

$$|||\Theta + \Gamma|||_1 = |||\Theta|||_1 + |||\Gamma|||_1 \qquad \text{for all } \Theta \in A(U,V) \text{ and } \Gamma \in B^{\perp}(U,V).$$

Main theorem

Estimator

$$\widehat{\theta}_{\lambda_n} \quad \in \quad \arg\min_{\theta \in \mathbb{R}^d} \big\{ \mathcal{L}_n(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \big\},$$

where $\mathcal{L}$ satisfies RSC(γ, τ) w.r.t regularizer $\mathcal{R}$.

Main theorem

Estimator

$$\widehat{\theta}_{\lambda_n} \quad \in \quad \arg\min_{\theta \in \mathbb{R}^d} \left\{ \mathcal{L}_n(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \right\},$$

where $\mathcal{L}$ satisfies RSC(γ, τ) w.r.t regularizer $\mathcal{R}$.

Theorem (Negahban, Ravikumar, W., & Yu, 2012)

Suppose that $\theta^* \in A$, and $\Psi^2(A)\tau_n^2 < 1$. Then for any regularization parameter $\lambda_n \geq 2\mathcal{R}^*\big(\nabla \mathcal{L}(\theta^*; Z_1^n)\big)$, any solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_\star^2 \precsim \frac{1}{\gamma^2(\mathcal{L})} \; \lambda_n^2 \; \Psi^2(A).$$

Quantities that control rates:

- curvature in RSC: γ_ℓ
- tolerance in RSC: τ
- dual norm of regularizer: $\mathcal{R}^*(v) := \sup_{\mathcal{R}(u) \leq 1} \langle v, u \rangle$.
- optimal subspace const.: $\Psi(A) = \sup_{\theta \in A \backslash \{0\}} \mathcal{R}(\theta)/\|\theta\|_\star$

Main theorem

Estimator

$$\widehat{\theta}_{\lambda_n} \quad \in \quad \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \mathcal{L}_n(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \Big\},$$

Theorem (Oracle version)

With $\lambda_n \geq 2\mathcal{R}^*\big(\nabla \mathcal{L}(\theta^*; Z_1^n)\big)$, any solution $\widehat{\theta}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_\star^2 \precsim \underbrace{\frac{(\lambda_n')^2}{\gamma^2} \Psi^2(A)}_{\textit{Estimation error}} \quad + \quad \underbrace{\frac{\lambda_n'}{\gamma} \mathcal{R}(\Pi_{A^\perp}(\theta^*))}_{\textit{Approximation error}}$$

where $\lambda' = \max\{\lambda, \tau\}$.

Quantities that control rates:

- curvature in RSC: γ_ℓ
- tolerance in RSC: τ
- dual norm of regularizer: $\mathcal{R}^*(v) := \sup_{\mathcal{R}(u) \leq 1} \langle v, u \rangle.$
- optimal subspace const.: $\Psi(A) = \sup_{\theta \in A \setminus \{0\}} \mathcal{R}(\theta) / \|\theta\|_\star$

Example: Group-structured regularizers

Many applications exhibit sparsity with more structure.....

- divide index set $\{1, 2, \ldots, d\}$ into groups $\mathcal{G} = \{G_1, G_2, \ldots, G_T\}$
- for parameters $\nu_i \in [1, \infty]$, define block-norm

$$\|\theta\|_{\nu,\mathcal{G}} := \sum_{t=1}^{T} \|\theta_{G_t}\|_{\nu_t}$$

- group/block Lasso program

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

- different versions studied by various authors
  (Wright et al., 2005; Tropp et al., 2006; Yuan & Li, 2006; Baraniuk, 2008; Obozinski et al., 2008; Zhao et al., 2008; Bach et al., 2009; Lounici et al., 2009)

Convergence rates for general group Lasso

Corollary

Say Θ^* is supported on group subset $\mathcal{S}_{\mathcal{G}}$, and X satisfies RSC. Then for regularization parameter

$$\lambda_n \geq 2 \max_{t=1,2,\ldots,T} \| \frac{X^T w}{n} \|_{\nu_t^*}, \qquad \text{where } \frac{1}{\nu_t^*} = 1 - \frac{1}{\nu_t},$$

any solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2 \leq \frac{2}{\gamma(\mathcal{L})} \Psi_\nu(S_{\mathcal{G}})\, \lambda_n, \qquad \text{where } \Psi_\nu(S_{\mathcal{G}}) = \sup_{\theta \in A(S_{\mathcal{G}}) \setminus \{0\}} \frac{\|\theta\|_{\nu,\mathcal{G}}}{\|\theta\|_2}.$$

Convergence rates for general group Lasso

Corollary

Say Θ^* is supported on group subset $\mathcal{S}_{\mathcal{G}}$, and X satisfies RSC. Then for regularization parameter

$$\lambda_n \geq 2 \max_{t=1,2,\ldots,T} \left\| \frac{X^T w}{n} \right\|_{\nu_t^*}, \qquad \textit{where } \frac{1}{\nu_t^*} = 1 - \frac{1}{\nu_t},$$

any solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2 \leq \frac{2}{\gamma(\mathcal{L})} \Psi_\nu(S_{\mathcal{G}}) \, \lambda_n, \qquad \textit{where } \Psi_\nu(S_{\mathcal{G}}) = \sup_{\theta \in A(S_{\mathcal{G}}) \setminus \{0\}} \frac{\|\theta\|_{\nu,\mathcal{G}}}{\|\theta\|_2}.$$

Some special cases with $m \equiv$ max. group size

1. ℓ_1/ℓ_2 regularization: Group norm with $\nu = 2$

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 = \mathcal{O}\Big(\frac{|\mathcal{S}_{\mathcal{G}}| \, m}{n} + \frac{|\mathcal{S}_{\mathcal{G}}| \log T}{n} \Big).$$

Convergence rates for general group Lasso

Corollary

Say Θ^* is supported on group subset $\mathcal{S}_{\mathcal{G}}$, and X satisfies RSC. Then for regularization parameter

$$\lambda_n \geq 2 \max_{t=1,2,\ldots,T} \left\| \frac{X^T w}{n} \right\|_{\nu_t^*}, \qquad \text{where } \frac{1}{\nu_t^*} = 1 - \frac{1}{\nu_t},$$

any solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2 \leq \frac{2}{\gamma(\mathcal{L})} \Psi_\nu(S_{\mathcal{G}})\, \lambda_n, \qquad \text{where } \Psi_\nu(S_{\mathcal{G}}) = \sup_{\theta \in A(S_{\mathcal{G}}) \setminus \{0\}} \frac{\|\theta\|_{\nu,\mathcal{G}}}{\|\theta\|_2}.$$

Some special cases with $m \equiv$ max. group size

1. ℓ_1/ℓ_∞ regularization: group norm with $\nu = \infty$

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2^2 = \mathcal{O}\Big(\frac{|\mathcal{S}_{\mathcal{G}}| m^2}{n} + \frac{|\mathcal{S}_{\mathcal{G}}|\, \log T}{n}\Big).$$

Is adaptive estimation possible?

Consider a group-sparse problem with:

- T groups in total
- each of size m
- $|\mathcal{S}_{\mathcal{G}}|$-active groups
- T active coefficients per group

Group Lasso will achieve

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{|\mathcal{S}_{\mathcal{G}}| m}{n} + \frac{|\mathcal{S}_{\mathcal{G}}| \log |\mathcal{G}|}{n}.$$

Lasso will achieve

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{|\mathcal{S}_{\mathcal{G}}|\, T\, \log(|\mathcal{G}| m)}{n}.$$

Is adaptive estimation possible?

Consider a group-sparse problem with:

- T groups in total
- each of size m
- $|\mathcal{S}_\mathcal{G}|$-active groups
- T active coefficients per group

Group Lasso will achieve

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{|\mathcal{S}_\mathcal{G}|m}{n} + \frac{|\mathcal{S}_\mathcal{G}| \log |\mathcal{G}|}{n}.$$

Lasso will achieve

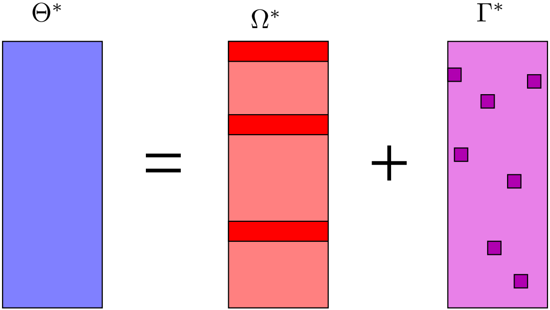
$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{|\mathcal{S}_\mathcal{G}|\, T \, \log(|\mathcal{G}|m)}{n}.$$

Natural question:

Can we design an estimator that optimally adapts to the degree of elementwise versus group sparsity?

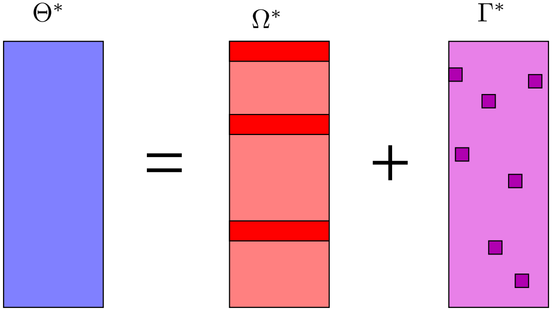
Answer: Overlap group Lasso

Represent Θ^* as a sum of row-sparse and element-wise sparse matrices.



Answer: Overlap group Lasso

Represent Θ^* as a sum of row-sparse and element-wise sparse matrices.

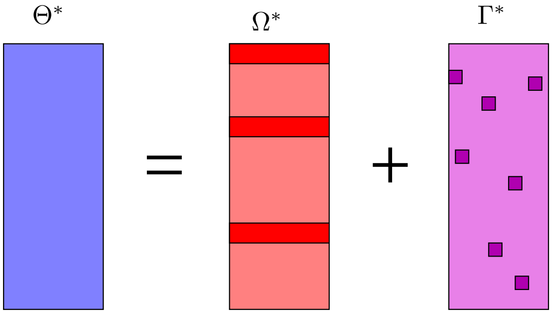


Define new norm on matrix space:

$$\mathcal{R}_\omega(\Theta) = \inf_{\Theta=\Omega+\Gamma} \Big\{ \omega \|\Omega\|_{1,2} + \|\Gamma\|_1 \Big\}.$$

Answer: Overlap group Lasso

Represent Θ^* as a sum of row-sparse and element-wise sparse matrices.



Define new norm on matrix space:

$$\mathcal{R}_\omega(\Theta) = \inf_{\Theta=\Omega+\Gamma} \Big\{ \omega \|\Omega\|_{1,2} + \|\Gamma\|_1 \Big\}.$$

Special case of the overlap group Lasso: (Obozinski et al., 2008; Jalali et al., 2011)

Example: Adaptivity with overlap group Lasso

Consider regularizer

$$\mathcal{R}_\omega(\Theta) = \inf_{\Theta=\Omega+\Gamma} \Big\{ \omega \|\Omega\|_{1,2} + \|\Gamma\|_1 \Big\}.$$

with

$$\omega = \frac{\sqrt{m} + \sqrt{\log |\mathcal{G}|}}{\sqrt{\log d}},$$

- $|\mathcal{G}|$ is number of groups
- m is max. group size
- d is number of predictors.

Example: Adaptivity with overlap group Lasso

Consider regularizer

$$\mathcal{R}_\omega(\Theta) = \inf_{\Theta=\Omega+\Gamma} \Big\{\omega\|\Omega\|_{1,2} + \|\Gamma\|_1\Big\}.$$

with

$$\omega = \frac{\sqrt{m} + \sqrt{\log|\mathcal{G}|}}{\sqrt{\log d}},$$

- $|\mathcal{G}|$ is number of groups
- m is max. group size
- d is number of predictors.

Corollary

Under RSC condition on loss function, suppose that Θ^* can be decomposed as a sum of an $|S_{elt}|$-elementwise sparse matrix and an $|\mathcal{S}_{\mathcal{G}}|$-groupwise sparse matrix (disjointly). Then for $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$, any optimal solution satisfies (w.h.p.)

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \precsim \sigma^2 \left\{\frac{|\mathcal{S}_{\mathcal{G}}|m}{n} + \frac{|\mathcal{S}_{\mathcal{G}}|\log|\mathcal{G}|}{n}\right\} + \sigma^2\left\{\frac{|S_{elt}|\log d}{n}\right\}.$$

Example: Low-rank matrices and nuclear norm

- low-rank matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ that is exactly (or approximately) low-rank
- noisy/partial observations of the form

$$y_i \quad = \quad \langle\langle X_i,\ \Theta^* \rangle\rangle + w_i,\ i = 1, \ldots, n, \quad w_i \quad \text{i.i.d. noise}$$

- estimate by solving semi-definite program (SDP):

$$\widehat{\Theta} \in \arg\min_{\Theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\langle X_i,\ \Theta \rangle\rangle)^2 + \lambda_n \underbrace{\sum_{j=1}^{\min\{p_1,p_2\}} \gamma_j(\Theta)}_{|||\Theta|||_1} \Big\}$$

Example: Low-rank matrices and nuclear norm

- low-rank matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ that is exactly (or approximately) low-rank
- noisy/partial observations of the form

$$y_i \quad = \quad \langle\!\langle X_i, \; \Theta^* \rangle\!\rangle + w_i, \; i = 1, \ldots, n, \quad w_i \quad \text{i.i.d. noise}$$

- estimate by solving semi-definite program (SDP):

$$\widehat{\Theta} \in \arg\min_{\Theta} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\!\langle X_i, \; \Theta \rangle\!\rangle)^2 + \lambda_n \underbrace{\sum_{j=1}^{\min\{p_1, p_2\}} \gamma_j(\Theta)}_{|\!|\!|\Theta|\!|\!|_1} \Big\}$$

- various applications:
  - matrix compressed sensing
  - matrix completion
  - rank-reduced multivariate regression (multi-task learning)
  - time-series modeling (vector autoregressions)
  - phase-retrieval problems

Rates for (near) low-rank estimation

For simplicity, consider matrix compressed sensing model: X_i are random sub-Gaussian projections).

For parameter $q \in [0, 1]$, set of near low-rank matrices:

$$\mathbb{B}_q(R_q) = \big\{\Theta^* \in \mathbb{R}^{p_1 \times p_2} \;\mid\; \sum_{j=1}^{\min\{p_1, p_2\}} |\sigma_j(\Theta^*)|^q \leq R_q\big\}.$$

Rates for (near) low-rank estimation

For simplicity, consider matrix compressed sensing model: X_i are random sub-Gaussian projections).

For parameter $q \in [0,1]$, set of near low-rank matrices:

$$\mathbb{B}_q(R_q) = \big\{\Theta^* \in \mathbb{R}^{p_1 \times p_2} \mid \sum_{j=1}^{\min\{p_1,p_2\}} |\sigma_j(\Theta^*)|^q \leq R_q\big\}.$$

Corollary (Negahban & W., 2011)

With regularization parameter $\lambda_n \geq 16\sigma\big(\sqrt{\frac{p_1}{n}} + \sqrt{\frac{p_2}{n}}\big)$, *we have w.h.p.*

$$\|\widehat{\Theta} - \Theta^*\|_F^2 \quad \leq \quad c_0 \, \frac{R_q}{\gamma(\mathcal{L})^2} \left(\frac{\sigma^2 \, (p_1 + p_2)}{n}\right)^{1-\frac{q}{2}}$$

Rates for (near) low-rank estimation

For parameter $q \in [0,1]$, set of near low-rank matrices:

$$\mathbb{B}_q(R_q) = \Big\{\Theta^* \in \mathbb{R}^{p_1 \times p_2} \;\Big|\; \sum_{j=1}^{\min\{p_1,p_2\}} |\sigma_j(\Theta^*)|^q \leq R_q\Big\}.$$

Corollary (Negahban & W., 2011)

With regularization parameter $\lambda_n \geq 16\sigma\big(\sqrt{\frac{p_1}{n}} + \sqrt{\frac{p_2}{n}}\big)$, *we have w.h.p.*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \quad \leq \quad c_0 \frac{R_q}{\gamma(\mathcal{L})^2} \left(\frac{\sigma^2\,(p_1+p_2)}{n}\right)^{1-\frac{q}{2}}$$

- for a rank r matrix M

$$|||M|||_1 = \sum_{j=1}^r \sigma_j(M) \leq \sqrt{r}\sqrt{\sum_{j=1}^r \sigma_j^2(M)} \;=\; \sqrt{r}\,|||M|||_F$$

- solve nuclear norm regularized program with $\lambda_n \geq \frac{2}{n}|||\sum_{i=1}^n w_i X_i|||_2$

Matrix completion

Random operator $\mathfrak{X} : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ with

$$\big[\mathfrak{X}(\Theta^*)\big]_i = d\,\Theta^*_{a(i)b(i)}$$

where $(a(i), b(i))$ is a matrix index sampled uniformly at random.

Matrix completion

Random operator $\mathfrak{X} : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ with

$$\big[\mathfrak{X}(\Theta^*)\big]_i = d\,\Theta^*_{a(i)b(i)}$$

where $(a(i), b(i))$ is a matrix index sampled uniformly at random.

Even in noiseless setting, model is unidentifiable:
Consider a rank one matrix:

$$\Theta^* = e_1 e_1^T \;=\; \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Matrix completion

Random operator $\mathfrak{X} : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ with

$$\big[\mathfrak{X}(\Theta^*)\big]_i = d\,\Theta^*_{a(i)b(i)}$$

where $(a(i), b(i))$ is a matrix index sampled uniformly at random.

Even in noiseless setting, model is unidentifiable:
Consider a rank one matrix:

$$\Theta^* = e_1 e_1^T \;=\; \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

Exact recovery based on eigen-incoherence involving leverage scores
(e.g., Recht & Candes, 2008; Gross, 2009)

A milder "spikiness" condition

Consider the "poisoned" low-rank matrix:

$$\Theta^* = \Gamma^* + \delta e_1 e_1^T \;=\; \Gamma^* + \delta \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

where Γ^* is rank $r-1$, all eigenectors perpendicular to e_1.

Excluded by eigen-incoherence for all $\delta > 0$.

A milder "spikiness" condition

Consider the "poisoned" low-rank matrix:

$$\Theta^* = \Gamma^* + \delta e_1 e_1^T \;=\; \Gamma^* + \delta \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

where Γ^* is rank $r-1$, all eigenectors perpendicular to e_1.

Excluded by eigen-incoherence for all $\delta > 0$.

Control by spikiness ratio:

$$1 \;\leq\; \frac{d\|\Theta^*\|_\infty}{|||\Theta^*|||_F} \;\leq\; d.$$

Spikiness constraints used in various papers: Oh et al., 2009; Negahban & W. 2010, Koltchinski et al., 2011.

Uniform law for matrix completion

Let $\mathfrak{X}_n : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ be rescaled matrix completion random operator

$$(\mathfrak{X}_n(\Theta))_i \mapsto d\, \Theta_{a(i),b(i)} \quad \text{where index } (a(i), b(i)) \text{ from uniform distribution.}$$

Define family of zero-mean random variables:

$$Z_n(\Theta) := \frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n} - \|\Theta\|_F^2, \qquad \text{for } \Theta \in \mathbb{R}^{d \times d}.$$

Uniform law for matrix completion

Let $\mathfrak{X}_n : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ be rescaled matrix completion random operator

$$(\mathfrak{X}_n(\Theta))_i \mapsto d\, \Theta_{a(i),b(i)} \quad \text{where index } (a(i), b(i)) \text{ from uniform distribution.}$$

Define family of zero-mean random variables:

$$Z_n(\Theta) := \frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n} - \|\Theta\|_F^2, \qquad \text{for } \Theta \in \mathbb{R}^{d \times d}.$$

Theorem (Negahban & W., 2010)

For random matrix completion operator $\mathfrak{X}_n$, there are universal positive constants (c_1, c_2) such that

$$\sup_{\Theta \in \mathbb{R}^{d \times d} \setminus \{0\}} Z_n(\Theta) \leq \underbrace{c_1\, d \|\Theta\|_\infty \, |\!|\!|\Theta|\!|\!|_{nuc} \sqrt{\frac{d \log d}{n}}}_{\text{“low-rank term”}} + \underbrace{c_2 \Big(d\|\Theta\|_\infty \sqrt{\frac{d \log d}{n}} \Big)^2}_{\text{“spikiness” term}}$$

with probability at least $1 - \exp(-d \log d)$.