High-dimensional statistics: Some progress and challenges ahead

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS
Winedale Workshop
October 2010

Joint work with: Alekh Agarwal, Arash Amini, Sahand Negahban
Pradeep Ravikumar, Bin Yu.

Introduction

- classical asymptotic theory: sample size $n \to +\infty$ with number of parameters p fixed

- modern applications in science and engineering:
  - large-scale problems: both p and n may be large (possibly $p \gg n$)
  - need for high-dimensional theory that allows $(n, p) \to +\infty$

Introduction

- classical asymptotic theory: sample size $n \to +\infty$ with number of parameters p fixed

- modern applications in science and engineering:
  - large-scale problems: both p and n may be large (possibly $p \gg n$)
  - need for high-dimensional theory that allows $(n, p) \to +\infty$

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

Introduction

- classical asymptotic theory: sample size $n \to +\infty$ with number of parameters p fixed
- modern applications in science and engineering:
  - large-scale problems: both p and n may be large (possibly $p \gg n$)
  - need for high-dimensional theory that allows $(n, p) \to +\infty$
- curses and blessings of high dimensionality
  - exponential explosions in computational complexity
  - statistical curses (sample complexity)
  - concentration of measure
- need for embedded low-dimensional structures
  - sparse vectors (compressed sensing)
  - structured/patterned matrices
  - (near) low-rank matrices
  - Markov random fields
  - manifold structure

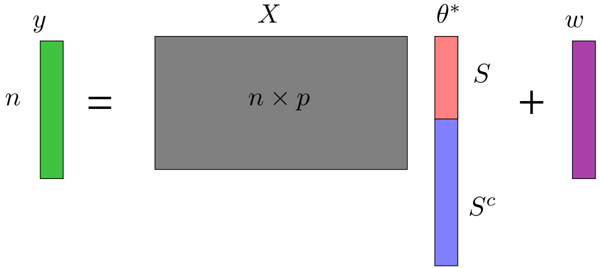
Loss functions and regularization

- **Models:** Indexed class of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$
- **Data:** samples $Z_1^n = (x_i, y_i)$, $i = 1, \ldots, n$ drawn from unknown $\mathbb{P}_{\theta^*}$
- **Estimation:** Minimize loss function plus regularization term:

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

- **Goal:** For given error norm $\|\cdot\|_\star$
  - want upper bounds on $\|\widehat{\theta} - \theta^*\|_\star$
  - non-asymptotic results allowing for $(n, p, s_1, s_2, \ldots) \to \infty$ where
    - $n \equiv$ sample size
    - $p \equiv$ dimension of parameter space Ω
    - $s_i \equiv$ structural parameters (e.g., sparsity, rank, graph degree)

Example: Sparse linear regression



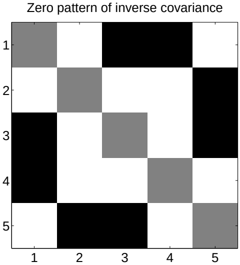
Set-up: noisy observations $y = X\theta^* + w$ with sparse θ^*

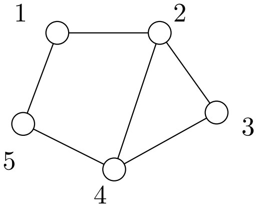
Estimator: Lasso program

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

<u>Some past work:</u> Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Efron et al., 2004; Meinshausen & Buhlmann, 2005; Candes & Tao, 2005; Donoho, 2005; Haupt & Nowak, 2005; Zhou & Yu, 2006; Zou, 2006; Koltchinskii, 2007; van de Geer, 2007; Bickel, Ritov & Tsybakov, 2008, Zhang, 2009 .....

Example: Structured (inverse) covariance matrices





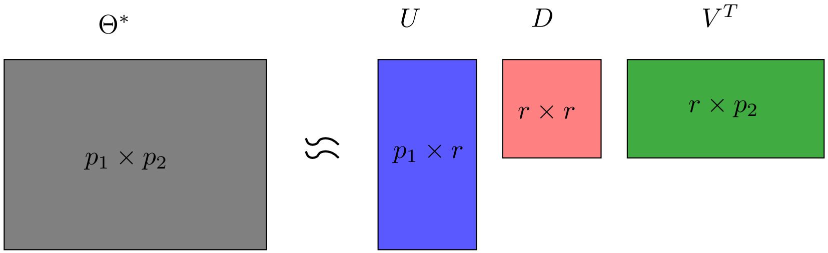
Set-up: Samples from random vector with sparse covariance Σ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{p \times p}$.

Estimator (for inverse covariance)

$$\widehat{\Theta} \in \arg\min_{\Theta} \Big\{ \langle\!\langle \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T, \; \Theta \rangle\!\rangle - \log\det(\Theta) + \lambda_n \sum_{b \in B} |\!|\!| \Theta_b |\!|\!|_F \Big\}$$

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

Example: Low-rank matrix approximation



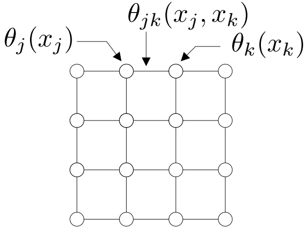
Set-up: Matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ with rank $r \ll \min\{p_1, p_2\}$.

Estimator:

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

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

Example: Discrete Markov random fields



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

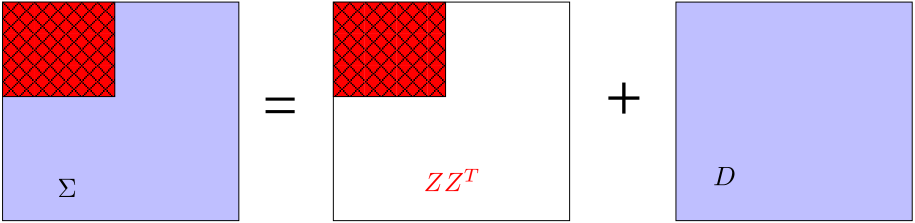
$$\mathbb{P}_\theta(x_1, \ldots, x_p) = \frac{1}{Z(\theta)} \exp\Big\{ \sum_{j \in V} \theta_j(x_j) + \sum_{(j,k) \in E} \theta_{jk}(x_j, x_k) \Big\}.$$

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

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

Motivation and outline

- a large number of high-dimensional models and associated results on regularized estimators
- is there a core set of ideas that underlie these analyses?

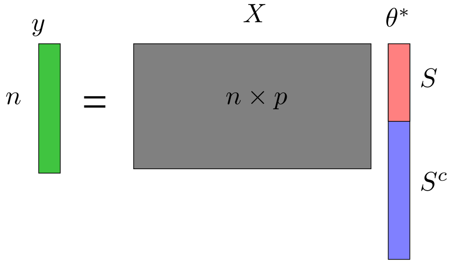
Motivation and outline

- a large number of high-dimensional models and associated results on regularized estimators
- is there a core set of ideas that underlie these analyses?

This tutorial:

1. Part I: Linear regression with sparsity constraints
   - Restricted nullspace and ℓ_1-minimization
   - A random matrix theory result
   - Restricted eigenvalues and Lasso
2. Part II: A more general theory
   - Decomposability of regularizers
   - Restricted strong convexity of loss function
   - A main theorem
   - Some consequences

Noiseless linear models and basis pursuit



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

<u>ℓ_0-optimization</u>

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

Computationally intractable
NP-hard

<u>ℓ_1-relaxation</u>

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

Linear program (easy to solve)
Basis pursuit relaxation

Restricted nullspace: necessary and sufficient

Definition

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

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

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

Restricted nullspace: necessary and sufficient

Definition

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

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

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

Proposition

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

Restricted nullspace: necessary and sufficient

Definition

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

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

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

Proof (sufficiency):

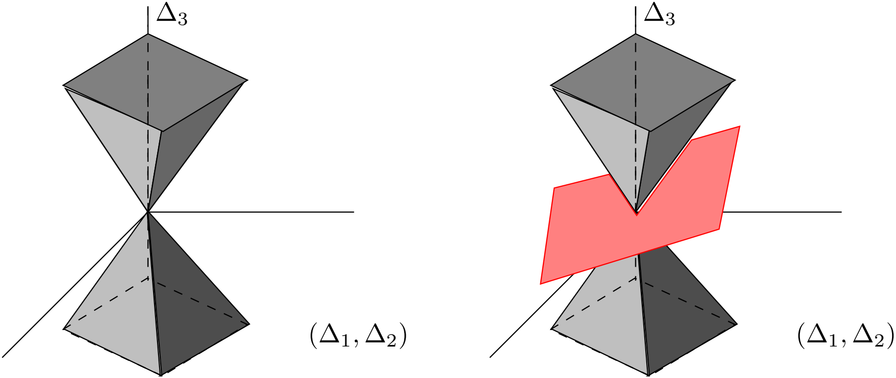
(1) Error vector $\widehat{\Delta} = \theta^* - \widehat{\theta}$ satisfies $X\widehat{\Delta} = 0$, and hence $\widehat{\Delta} \in \mathbb{N}(X)$.

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

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

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

Illustration of restricted nullspace property



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

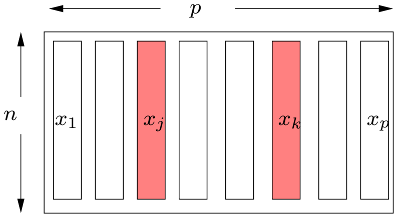
$$\mathbb{C}(S; 1) := \Big\{(\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3 \mid |\Delta_1| + |\Delta_2| \leq |\Delta_3|\Big\}.$$

Some sufficient conditions

How to verify RN property for a given sparsity s?

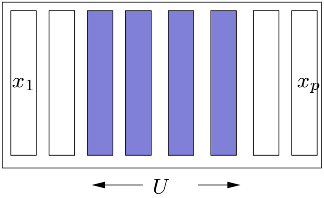
1. Elementwise incoherence condition (Donoho & Xuo, 2001; Feuer & Nem., 2003)

$$\max_{j,k=1,\ldots,p} \left| \frac{\langle x_j,\, x_k \rangle}{n} - \mathbb{I}[j=k] \right| \leq \frac{\delta_1}{s}$$



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

$$\max_{|U| \leq 2s} \left\| \frac{X_U^T X_U}{n} - I_{|U| \times |U|} \right\|_2 \leq \delta_{2s}.$$

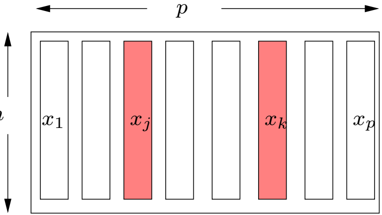


Some sufficient conditions

How to verify RN property for a given sparsity s?

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

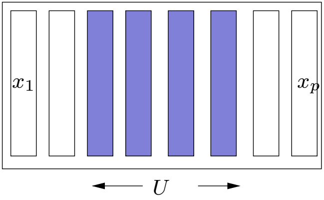
$$\max_{j,k=1,\ldots,p} \left| \frac{\langle x_j,\, x_k \rangle}{n} - \mathbb{I}\left[j = k\right] \right| \leq \frac{\delta_1}{s}$$



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

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

$$\max_{|U| \leq 2s} \left\| \left\| \frac{X_U^T X_U}{n} - I_{|U| \times |U|} \right\| \right\|_2 \leq \delta_{2s}.$$



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

Violating matrix incoherence (elementwise/RIP)

Important:

Incoherence/RIP conditions imply RN, but are far from necessary.
Very easy to violate them.....

Violating matrix incoherence (elementwise/RIP)

Form random design matrix

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

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

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

Violating matrix incoherence (elementwise/RIP)

Form random design matrix

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

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

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

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

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

Violating matrix incoherence (elementwise/RIP)

Form random design matrix

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

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

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

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

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

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

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

Direct result for restricted nullspace/eigenvalues

Theorem (Raskutti, W., & Yu, 2009)

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

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

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

Direct result for restricted nullspace/eigenvalues

Theorem (Raskutti, W., & Yu, 2009)

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

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

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

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

Easy verification of restricted nullspace

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

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

- applying previous result:

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

Easy verification of restricted nullspace

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

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

- applying previous result:

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

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

Easy verification of restricted nullspace

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

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

- applying previous result:

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

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

Definition

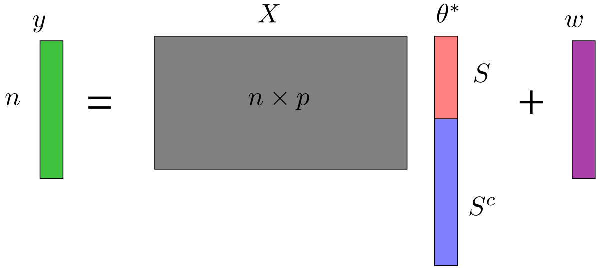
A design matrix $X \in \mathbb{R}^{n\times p}$ satisfies the *restricted eigenvalue* (RE) condition over S (denote RE(S)) with parameters $\alpha \geq 1$ and $\gamma > 0$ if

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

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

Lasso and restricted eigenvalues

Turning to noisy observations...



Estimator: Lasso program

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

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

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

---

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

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

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

---

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

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

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

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

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

---

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

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

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

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

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

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

Lasso bounds: Four simple steps

Let's analyze constrained version:

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

---

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

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

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

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

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

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

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

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

Lasso error bounds for different models

Proposition

Suppose that

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

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

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

Lasso error bounds for different models

Proposition

Suppose that

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

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

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

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

Lasso error bounds for different models

Proposition

Suppose that

- vector θ^* has support S, with cardinality s, and
- design matrix X satisfies $\mathrm{RE}(S)$ 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 $w_i \sim N(0, \sigma^2)$

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

Part II: A more general theory

Recap: Thus far.....

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

Part II: A more general theory

Recap: Thus far.....

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

The big picture:

Lots of other estimators with same basic form:

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

Part II: A more general theory

Recap: Thus far.....

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

The big picture:

Lots of other estimators with same basic form:

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

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?

Part II: A more general theory

Recap: Thus far.....

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

The big picture:

Lots of other estimators with same basic form:

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

Question: Is there a common set of underlying principles?

Answer: Yes, two essential ingredients.

1. Decomposability of the regularizer
2. Restricted strong convexity of the objective function

Decomposable regularizers

Definition

A norm-based regularizer is decomposable with respect to a pair of subspaces $A \subseteq B$ if

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

$\alpha \in A$	Model/ideal vector
$\beta \in B^\perp$	Perturbation away from ideal

Decomposable regularizers

Definition

A norm-based regularizer is decomposable with respect to a pair of subspaces $A \subseteq B$ if

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

$\alpha \in A$	Model/ideal vector
$\beta \in B^{\perp}$	Perturbation away from ideal

Intuition:

- By triangle inequality, we always have

$$r(\alpha + \beta) \leq r(\alpha) + r(\beta).$$

- "Tough love": Decomposable regularizers penalize perturbation as much as possible.

Examples of decomposable regularizers

- Sparse vectors and ℓ_1-regularization:
  - for each subset $S \subset \{1, \ldots, p\}$, define subspace pairs

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

  - decomposability of ℓ_1-norm:

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

Examples of decomposable regularizers

- Sparse vectors and ℓ_1-regularization:
  - for each subset $S \subset \{1, \ldots, p\}$, define subspace pairs

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

  - decomposability of ℓ_1-norm:

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

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

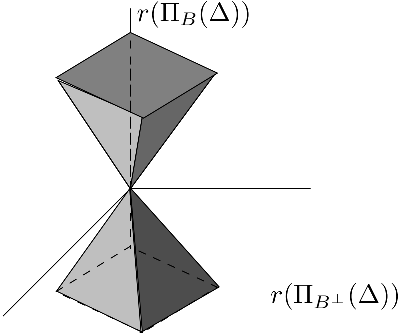
$$\begin{aligned} A(U, V) &:= \big\{\Theta \in \mathbb{R}^{p_1 \times p_2} \mid \text{row}(\Theta) \subseteq V, \; \text{col}(\Theta) \subseteq U\big\} \\ B^\perp(U, V) &:= \big\{\Gamma \in \mathbb{R}^{p_1 \times p_2} \mid \text{row}(\Gamma) \subseteq V^\perp, \; \text{col}(\Gamma) \subseteq U^\perp\big\}.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

    By construction, $\langle\!\langle \Theta, \; \Gamma \rangle\!\rangle = 0$ for all $\Theta \in A(U, V)$ and $\Gamma \in (U, V)$.

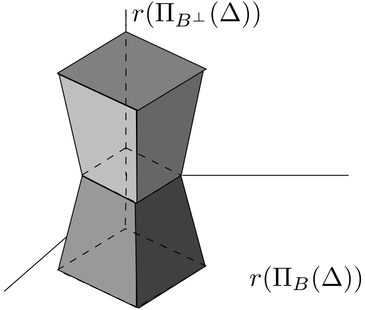
  - decomposability of nuclear norm $|\!|\!|\Theta|\!|\!|_1 = \sum_{j=1}^{\min\{p_1, p_2\}} \sigma_j(\Theta)$:

$$|\!|\!|\Theta + \Gamma|\!|\!|_1 \; = \; |\!|\!|\Theta|\!|\!|_1 + |\!|\!|\Gamma|\!|\!|_1 \qquad \forall \quad \Theta \in A(U, V), \Gamma \in B^\perp(U, V)$$

Significance of decomposability



(a) $\mathbb{C}$ for exact model (cone)



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

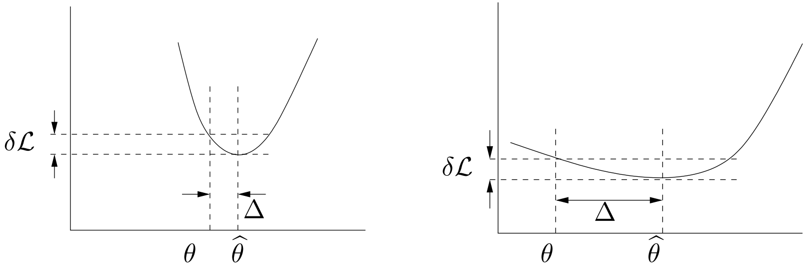
Lemma

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

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

Role of curvature

1. Curvature controls difficulty of estimation:

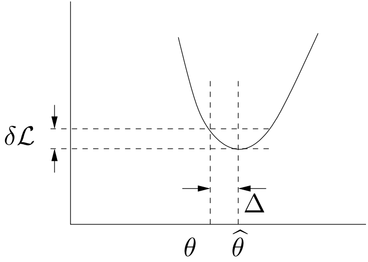


High curvature: easy to estimate

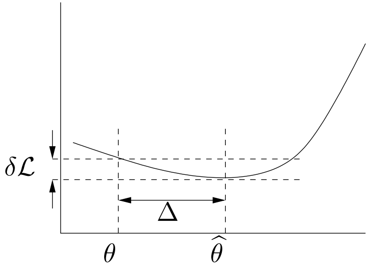
(b) Low curvature: harder

Role of curvature

1. Curvature controls difficulty of estimation:



High curvature: easy to estimate



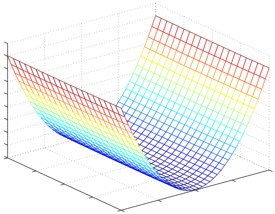
(b) Low curvature: harder

2. Curvature captured by strong convexity constant $c > 0$

$$\underbrace{\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*),\, \Delta \rangle}_{\delta\mathcal{L}(\Delta, \theta^*)} \geq c\|\Delta\|_\star^2$$

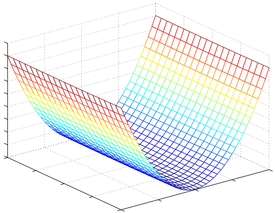
for all Δ in a neighborhood of θ^*.

Restricted strong convexity



For $p \gg n$, loss is flat in at least $p - n$ directions.

Restricted strong convexity



For $p \gg n$, loss is flat in at least $p - n$ directions.

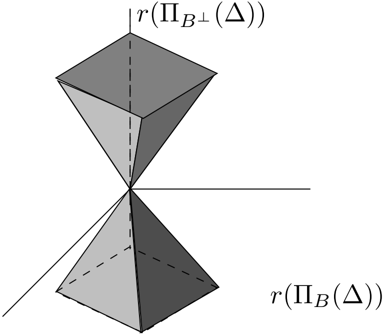
Definition

Loss function $\mathcal{L}(\theta) = \mathcal{L}(\theta; Z_1^n)$ satisfies restricted strong convexity (RSC) over a set $\mathbb{K}$

$$\underbrace{\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*)}_{\text{Excess loss}} - \langle \underbrace{\nabla \mathcal{L}(\theta^*)}_{\substack{\text{score}\\\text{function}}}, \Delta \rangle \quad \geq \quad \gamma(\mathcal{L}) \|\Delta\|_\star^2 \qquad \text{for all } \Delta \in \mathbb{K}.$$

When $\mathbb{K} = \mathbb{C}(S)$, natural generalization of restricted nullspace/eigenvalue conditions.

What sets to use for restricted strong convexity?

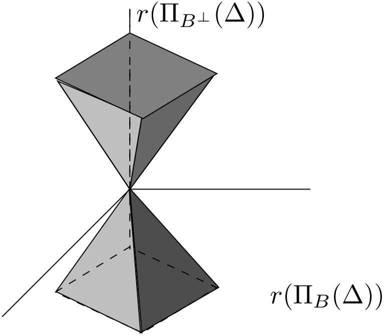


(a) Exact case: $\theta^* \in A$

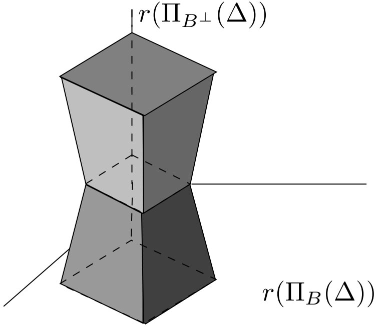
- For exact case, RSC can hold over

$$\mathbb{C}(A, B; \theta^*) := \left\{ \Delta \in \Omega \mid r(\Pi_{B^\perp}(\Delta)) \leq 3r(\Pi_B(\Delta)) + \underbrace{4r(\Pi_{A^\perp}(\theta^*))}_{\text{Zero when } \theta^* \in A} \right\}.$$

What sets to use for restricted strong convexity?



(a) Exact case: $\theta^* \in A$



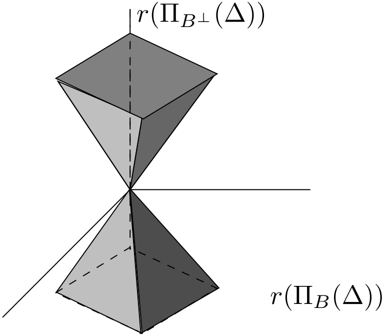
(b) Approximate case: $\theta^* \notin A$

- For exact case, RSC can hold over

$$\mathbb{C}(A, B; \theta^*) := \big\{\Delta \in \Omega \mid r(\Pi_{B^\perp}(\Delta)) \leq 3r(\Pi_B(\Delta)) + \underbrace{4r(\Pi_{A^\perp}(\theta^*))}_{\text{Zero when } \theta^* \in A}\big\}.$$

- For approximate case, RSC never holds over $\mathbb{C}(A, B; \theta^*)$.

What sets to use for restricted strong convexity?



(a) Exact case: $\theta^* \in A$ (b) Approximate case: $\theta^* \notin A$

- For approximate case, $\mathbb{C}$ is not a cone:

$$\mathbb{C}(A, B; \theta^*) := \Big\{ \Delta \in \Omega \;\Big|\; r(\Pi_{B^\perp}(\Delta)) \leq 3 r(\Pi_B(\Delta)) + \underbrace{4 r(\Pi_{A^\perp}(\theta^*))}_{\text{Non-zero when } \theta^* \notin A} \Big\}$$

- Need to intersect with a ball of $\|\cdot\|_\star$ radius δ

$$\mathbb{K}(\delta, A, B; \theta^*) := \mathbb{C}(A, B; \theta^*) \;\cap\; \big\{ \Delta \in \mathbb{R}^p \;\big|\; \|\Delta\|_\star = \delta \big\}.$$

Main theorem

Estimator $\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n r(\theta) \right\}$.
Decomposable across subspace pair $A \subseteq B$, where A represents model constraints.

Theorem (Negahban, Ravikumar, W., & Yu, 2009)

Consider the regularized problem for strictly positive $\lambda_n \geq 2 r^*(\nabla \mathcal{L}(\theta^*; Z_1^n))$. If $\theta^* \in A$ and RSC holds over $\mathbb{C}(A, B; \theta^*)$, then any solution $\widehat{\theta}_{\lambda_n}$ satisfies

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|_2 \leq \frac{1}{\gamma(\mathcal{L})} \Psi(B)\, \lambda_n.$$

Quantities that control rates:

- restricted strong convexity parameter: $\gamma(\mathcal{L})$
- dual norm of regularizer: $r^*(v) := \sup_{r(u)=1} \langle v, u \rangle$.
- subspace const.: $\Psi(B) = \sup_{\theta \in B \setminus \{0\}} r(\theta) / \|\theta\|_\star$

Main theorem

Estimator $\widehat{\theta} \quad \in \quad \arg\min_{\theta \in \mathbb{R}^p} \big\{\mathcal{L}(\theta; Z_1^n) + \lambda_n r(\theta)\big\}$.
Decomposable across subspace pair $A \subseteq B$, where A represents model constraints.

Theorem (Negahban, Ravikumar, W., & Yu, 2009)

Consider the regularized problem for strictly positive $\lambda_n \geq 2r^*(\nabla \mathcal{L}(\theta^*; Z_1^n))$. *Define the critical tolerance*

$$\delta_n := \inf_{\delta > 0} \Big\{ \delta \;\mid\; \delta \geq \underbrace{\frac{2\lambda_n}{\gamma(\mathcal{L})} \Psi(B)}_{\mathcal{E}_{\text{err}}} + \underbrace{\sqrt{\frac{2\lambda_n r(\Pi_{A^\perp}(\theta^*))}{\gamma(\mathcal{L})}}}_{\mathcal{E}_{\text{app}}} \;\; \textit{and } \text{RSC } \textit{over } \mathbb{K}(\delta; A, B) \Big\}.$$

Then any solution $\widehat{\theta}_{\lambda_n}$ *satisfies the bound* $\|\widehat{\theta} - \theta^*\|_\star \leq \delta_n$.

Quantities that control rates:

- restricted strong convexity parameter: $\gamma(\mathcal{L})$
- dual norm of regularizer: $r^*(v) := \sup_{r(u)=1} \langle v, \, u \rangle$.
- subspace const.: $\Psi(B) = \sup_{\theta \in B \setminus \{0\}} r(\theta)/\|\theta\|_\star$

Example: Linear regression (exact sparsity)

- Lasso program: $\min_{\theta \in \mathbb{R}^p} \left\{ \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}$
- RSC corresponds to lower bound on restricted eigenvalues of $X^T X \in \mathbb{R}^{p \times p}$
- for a k-sparse vector, we have $\|\theta\|_1 \leq \sqrt{k}\, \|\theta\|_2$.

Corollary

Suppose that true parameter θ^* is exactly k-sparse. Under RSC and with $\lambda_n \geq 2\|\frac{X^T w}{n}\|_\infty$, then any Lasso solution satisfies $\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2}{\gamma(\mathcal{L})} \sqrt{k}\, \lambda_n$.

Example: Linear regression (exact sparsity)

- Lasso program: $\min_{\theta \in \mathbb{R}^p} \left\{ \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}$
- RSC corresponds to lower bound on restricted eigenvalues of $X^T X \in \mathbb{R}^{p \times p}$
- for a k-sparse vector, we have $\|\theta\|_1 \leq \sqrt{k}\, \|\theta\|_2$.

Corollary

Suppose that true parameter θ^* is exactly k-sparse. Under RSC and with $\lambda_n \geq 2\|\frac{X^T w}{n}\|_\infty$, then any Lasso solution satisfies $\|\widehat{\theta} - \theta^*\|_2 \leq \frac{2}{\gamma(\mathcal{L})} \sqrt{k}\, \lambda_n$.

Some stochastic instances: recover known results

- Compressed sensing: $X_{ij} \sim N(0,1)$ and bounded noise $\|w\|_2 \leq \sigma\sqrt{n}$
- Deterministic design: X with bounded columns and $w_i \sim N(0, \sigma^2)$

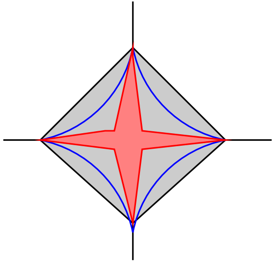
$$\|\frac{X^T w}{n}\|_\infty \leq \sqrt{\frac{3\sigma^2 \log p}{n}} \quad \text{w.h.p.} \implies \|\widehat{\theta} - \theta^*\|_2 \leq \frac{4\sigma}{\gamma(\mathcal{L})} \sqrt{\frac{k \log p}{n}}.$$

(e.g., Candes & Tao, 2007; Huang & Zhang, 2008; Meinshausen & Yu, 2008; Bickel et al., 2008)

Example: Linear regression (weak sparsity)

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

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



Corollary

For $\theta^* \in \mathbb{B}_q(R_q)$, any Lasso solution satisfies (w.h.p.)

$$\|\widehat{\theta} - \theta^*\|_2^2 \quad = \quad \mathcal{O}\Big[\sigma^2 R_q \Big(\frac{\log p}{n}\Big)^{1-q/2}\Big].$$

- rate known to be minimax optimal (Raskutti, W. & Yu, 2009)

Example: Generalized linear models (GLM)

- not all observation procsses are linear!
- generalized linear model linking covariates $x \in \mathbb{R}^p$ to output $y \in \mathcal{Y}$:

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

- Examples:
  - Ordinary linear observations: $\Phi(u) = u^2/2$
  - Bernoulli ($y \in \{-1, +1\}$): $\Phi(u) = \log(1 + \exp(u))$.
  - Poisson: $\Phi(u) = \exp(u)$.

Example: Generalized linear models (GLM)

- not all observation procsses are linear!
- generalized linear model linking covariates $x \in \mathbb{R}^p$ to output $y \in \mathcal{Y}$:

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

- Examples:
  - Ordinary linear observations: $\Phi(u) = u^2/2$
  - Bernoulli ($y \in \{-1, +1\}$): $\Phi(u) = \log(1 + \exp(u))$.
  - Poisson: $\Phi(u) = \exp(u)$.

Theorem (Negahban, Ravikumar, W. & Yu, 2010)

There exist constants (κ_1, κ_2), depending only on $(\psi, \mathrm{cov}(x))$, such that

$$\underbrace{\delta\mathcal{L}(\Delta, \theta^*;)}_{Taylor\ err} \geq \kappa_1 \|\Delta\|_2 \left\{ \|\Delta\|_2 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1 \right\} \qquad \textit{for all } \|\Delta\|_2 \leq 1$$

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

van de Geer (2008), Tewari et al. (2010): some related results.

Example: Group-structured regularizers

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

- divide index set $\{1, 2, \ldots, p\}$ 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

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

- 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)

Convergence rates for general group Lasso

Corollary

Say Θ^* *is supported on* $s_{\mathcal{G}}$ *groups, 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*

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

Convergence rates for general group Lasso

Corollary

Say Θ^* is supported on $s_{\mathcal{G}}$ groups, 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_2 regularization: Group norm with $\nu = 2$

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

This rate is minimax-optimal. (Raskutti, W. & Yu, 2010)

Convergence rates for general group Lasso

Corollary

Say Θ^* is supported on $s_{\mathcal{G}}$ groups, 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 $\nu = \infty$

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

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):

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

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):

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

- studied in past work (Fazel, 2001; Srebro et al., 2004; Bach, 2008)
- observations based on random projection (Recht, Fazel & Parillo, 2007)
- work on matrix completion (Srebro, 2004; Candes & Recht, 2008; Recht, 2009; Negahban & W., 2010)
- other work on general noisy observation models (Rohde & Tsybakov, 2009; Negahban & W., 2009)

Rates for (near) low-rank estimation

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

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

Corollary (Negahban & W., 2009)

Under RSC condition, 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:

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

Corollary (Negahban & W., 2009)

Under RSC condition, 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 $\lambda_n \geq \frac{2}{n}\|\sum_{i=1}^n w_i X_i\|_2$

Restricted strong convexity and nuclear norm

- observations $\{y_i = \langle\langle X_i,\ \Theta^*\rangle\rangle + w_i, i = 1, \ldots, n\}$ define *observation operator*

$$\mathfrak{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n, \qquad [\mathfrak{X}(\Delta)]_i = \langle\langle X_i,\ \Delta\rangle\rangle.$$

- restricted strong convexity for quadratic loss: $\frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{n}} \geq \gamma \, |\!|\!|\Delta|\!|\!|_F$ for all matrices $\Delta \in \mathbb{R}^{p_1 \times p_2}$ in

$$\mathbb{K} = \{|\!|\!|\Delta|\!|\!|_F = \delta\} \cap \big\{\Delta \mid \ |\!|\!|\Pi_{B^\perp}(\Delta)|\!|\!|_1 \leq 3|\!|\!|\Pi_B(\Delta)|\!|\!|_1 + 4|\!|\!|\Pi_{A^\perp}(\Theta^*)|\!|\!|_1\big\}$$

- let's consider this condition for standard random Gaussian matrices X_i

Restricted strong convexity and nuclear norm

- observations $\{y_i = \langle\langle X_i,\ \Theta^*\rangle\rangle + w_i, i = 1,\ldots,n\}$ define *observation operator*

$$\mathfrak{X} : \mathbb{R}^{p_1 \times p_2} \to \mathbb{R}^n, \qquad [\mathfrak{X}(\Delta)]_i = \langle\langle X_i,\ \Delta\rangle\rangle.$$

- restricted strong convexity for quadratic loss: $\frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{n}} \geq \gamma \, |||\Delta|||_F$ for all matrices $\Delta \in \mathbb{R}^{p_1 \times p_2}$ in

$$\mathbb{K} = \{|||\Delta|||_F = \delta\} \cap \big\{\Delta \mid \ |||\Pi_{B^\perp}(\Delta)|||_1 \leq 3|||\Pi_B(\Delta)|||_1 + 4|||\Pi_{A^\perp}(\Theta^*)|||_1\big\}$$

- let's consider this condition for standard random Gaussian matrices X_i

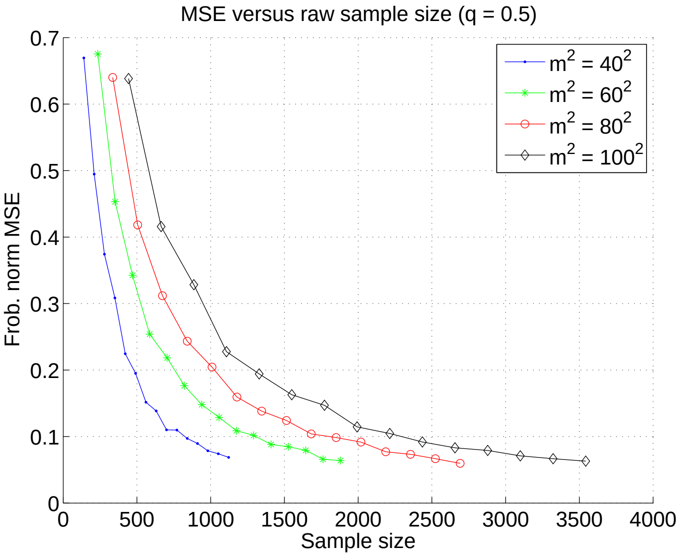
Proposition (Negahban & W., 2009)

Suppose that $X_i \in \mathbb{R}^{p_1 \times p_2}$ are i.i.d. random Gaussian matrices. Then

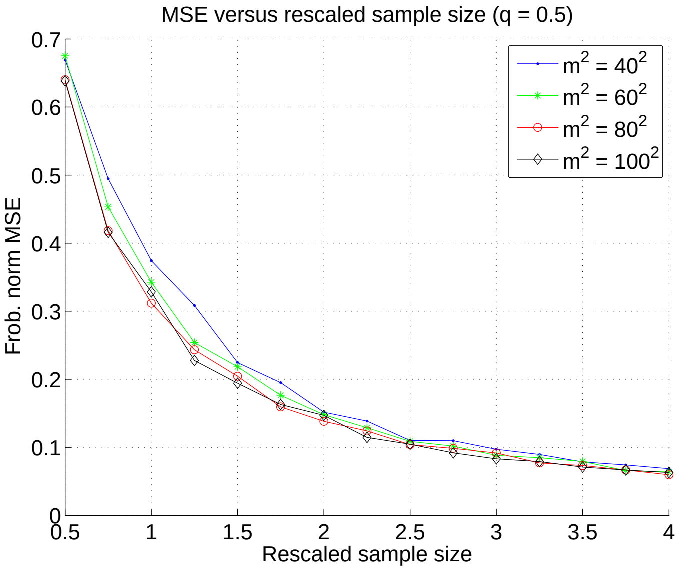
$$\frac{\|\mathfrak{X}(\Delta)\|_2}{\sqrt{n}} \geq \frac{1}{4}|||\Delta|||_F - \big(\sqrt{\frac{p_1}{n}} + \sqrt{\frac{p_2}{n}}\big)\ |||\Delta|||_1 \qquad \text{for all } \Delta \in \mathbb{R}^{p_1 \times p_2}$$

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

Results for noisy matrix completion (unrescaled)



Results for noisy matrix completion (rescaled)



Summary

- convergence rates for high-dimensional estimators
  - decomposability of regularizer r
  - restricted strong convexity of loss functions
- actual rates determined by:
  - noise measured in dual function r^*
  - subspace constant Ψ in moving from r to error norm $\|\cdot\|_\star$
  - restricted strong convexity constant
- recovered some known results as corollaries:
  - Lasso with hard sparsity
  - multivariate group Lasso
  - inverse covariance matrix estimation via log-determinant
- derived some new results on ℓ_2 or Frobenius norm:
  - models with weak sparsity
  - log-linear models with weak/exact sparsity
  - low-rank matrix estimation
  - other models? other error metrics?

Some papers (www.eecs.berkeley.edu/~wainwrig)

1. S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu (2010). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*. arxiv.org/abs/1010.2731.
2. S. Negahban and M. J. Wainwright (2009). Estimation rates of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, arxiv.org/abs/0912.5100.
3. S. Negahban and M. J. Wainwright (2010). Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. To appear in *Journal of Machine Learning Research*. arxiv.org/abs/0112.5100.
4. G. Raskutti, M. J. Wainwright and B. Yu (2011) Minimax rates for linear regression over ℓ_q-balls. *IEEE Trans. Information Theory*, arxiv.org/abs/arXiv:0910.2042.
5. G. Raskutti, M. J. Wainwright and B. Yu (2010). Restricted nullspace and eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*.