High-dimensional statistics: Some progress and challenges ahead

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

University College, London Master Class: Lecture 2

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

High-level overview

Last lecture: least-squares loss and ℓ_1-regularization.

The big picture:

Lots of other estimators with same basic form:

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

High-level overview

Last lecture: least-squares loss and ℓ_1-regularization.

The big picture:

Lots of other estimators with same basic form:

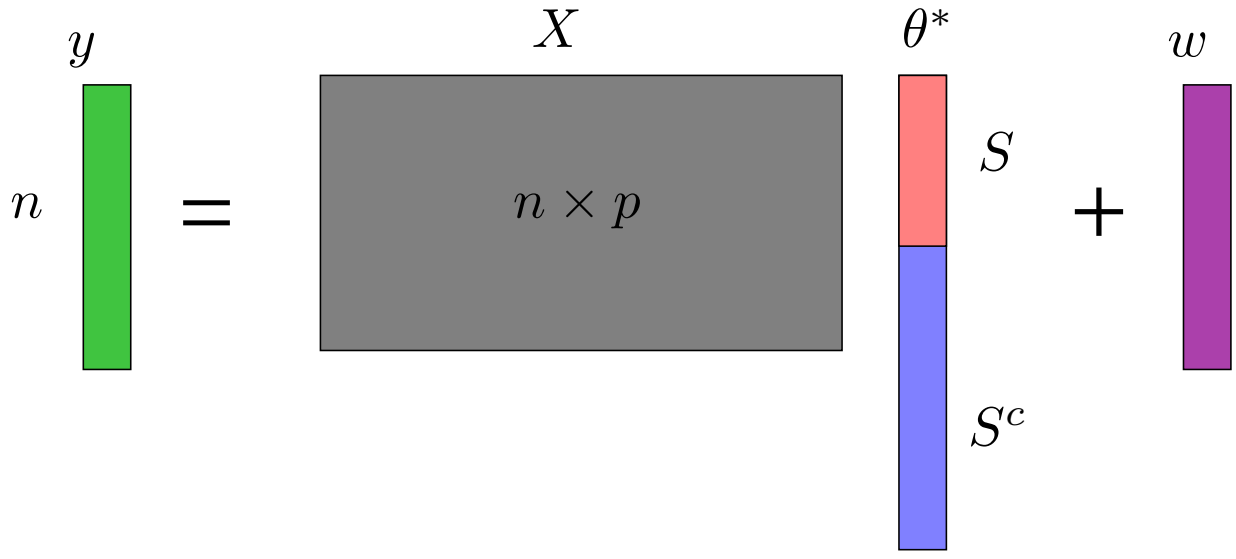
$$\underbrace{\widehat{\theta}_{\lambda_n}}_{\text{Estimate}} \in \arg\min_{\theta \in \Omega} \Big\{ \underbrace{\mathcal{L}(\theta; Z_1^n)}_{\text{Loss function}} + \lambda_n \underbrace{\mathcal{R}(\theta)}_{\text{Regularizer}} \Big\}.$$

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

Question:

Is there a common set of underlying principles?

Last lecture: Sparse linear regression

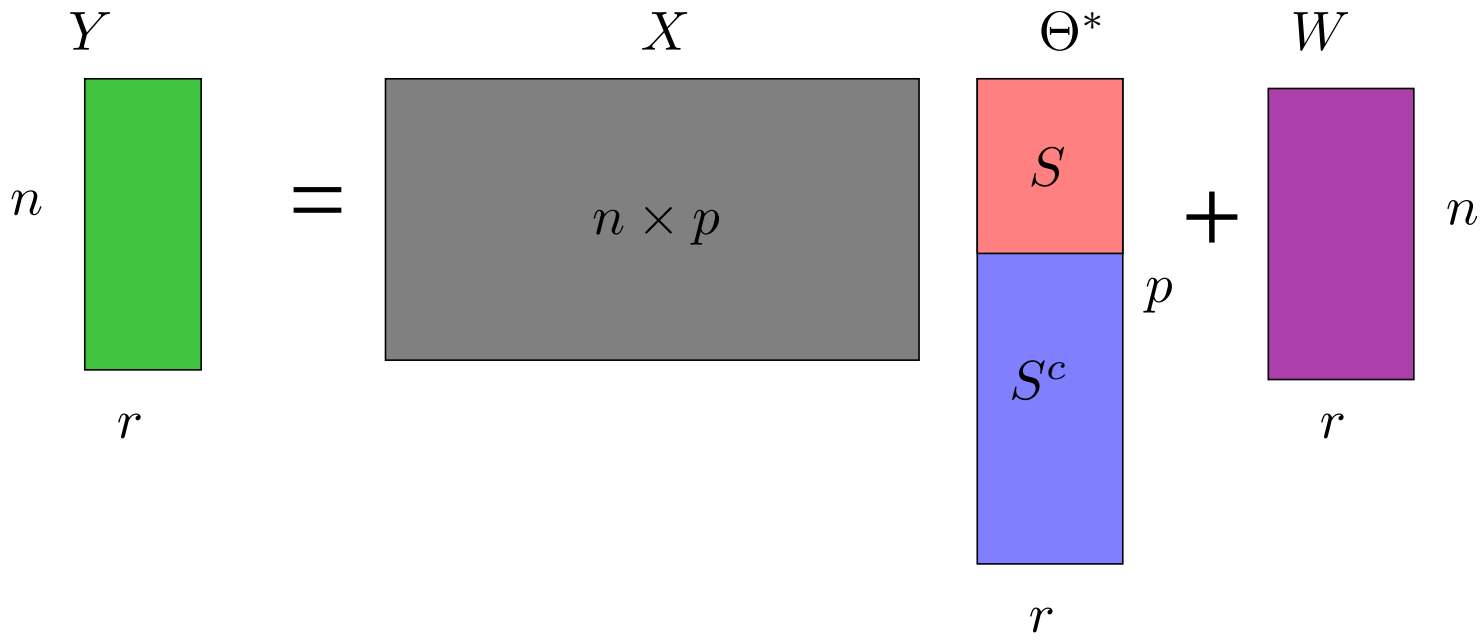


Set-up: noisy observations $y = X\theta^* + w$ with sparse θ^*

Estimator: Lasso program

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

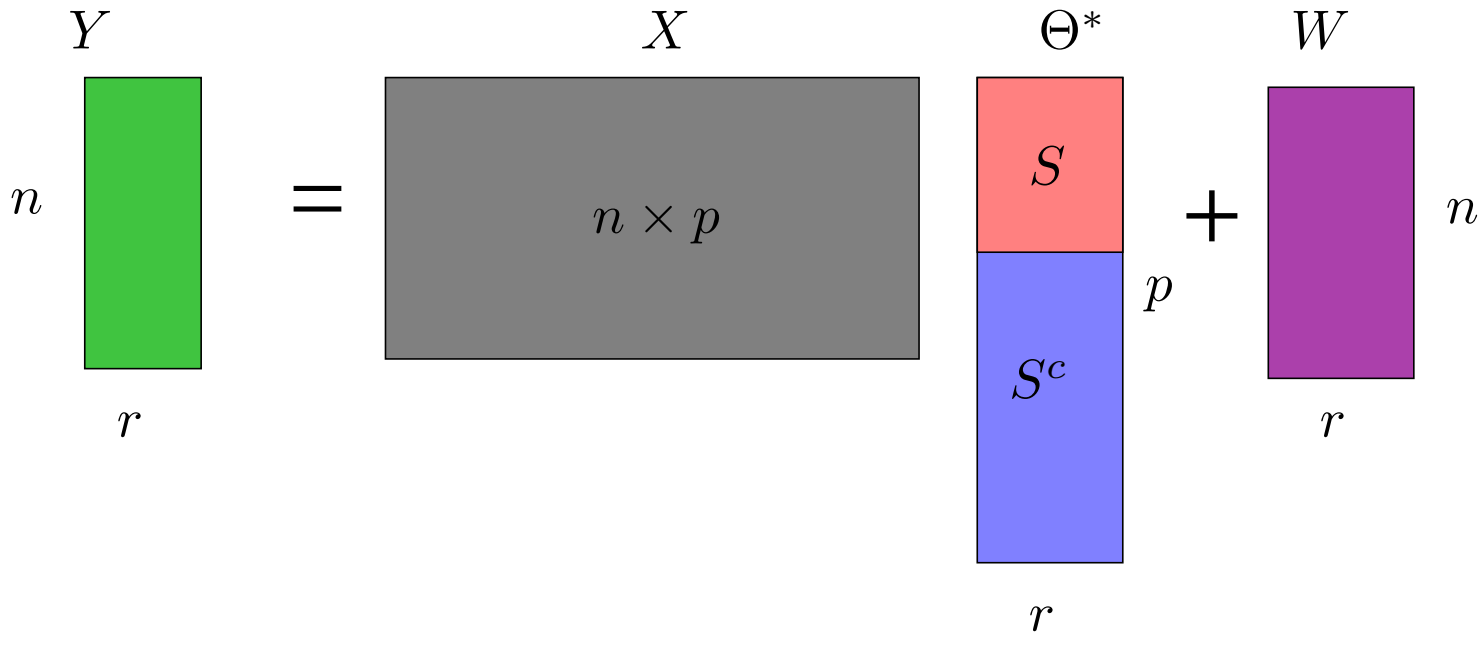
Block-structured extension



Signal Θ^* is a $p \times r$ matrix: partitioned into **non-zero rows** S and **zero rows** S^c

Various applications: multiple-view imaging, gene array prediction, graphical model fitting.

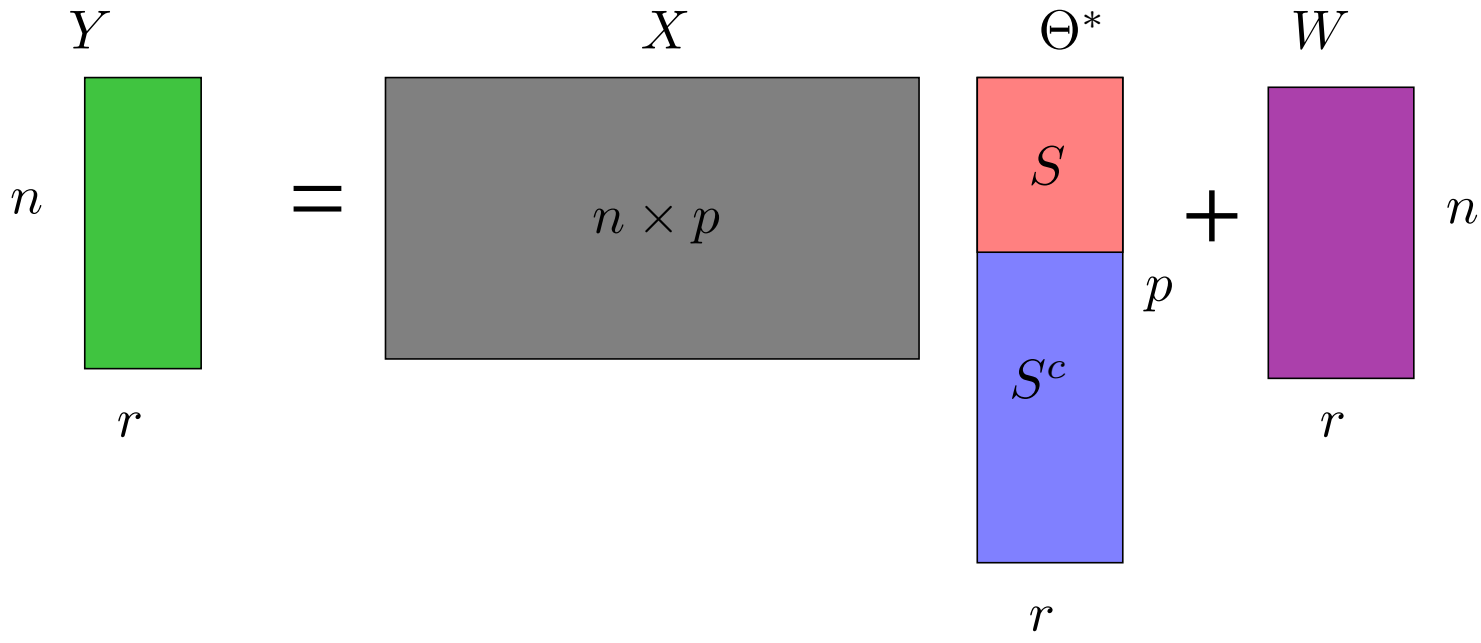
Block-structured extension



Row-wise ℓ_1/ℓ_2-norm

$$|||\Theta|||_{1,2} = \sum_{j=1}^{p} \|\Theta_j\|_2$$

Block-structured extension



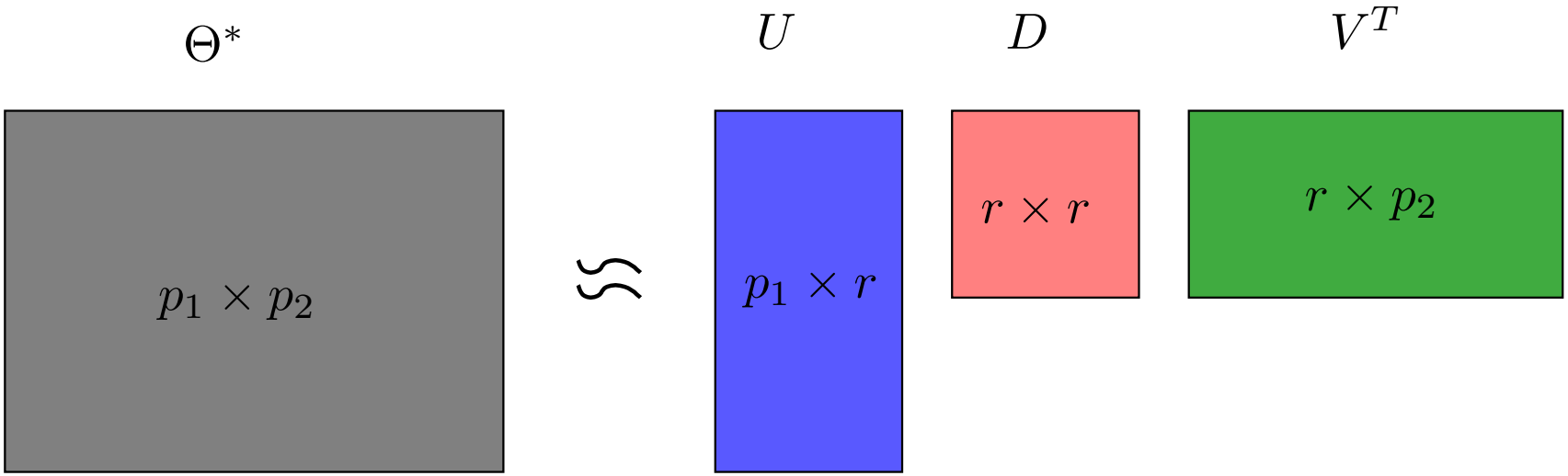
Row-wise ℓ_1/ℓ_2-norm

$$|||\Theta|||_{1,2} = \sum_{j=1}^{p} \|\Theta_j\|_2$$

More complicated group structure: (Obozinski et al., 2009)

$$|||\Theta^*|||_{\mathcal{G}} = \sum_{g \in \mathcal{G}} \|\Theta_g\|_2$$

Example: Low-rank matrix approximation



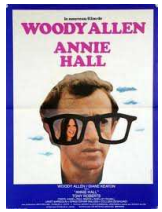
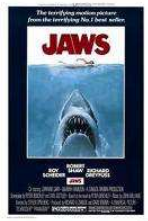
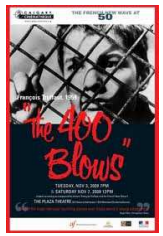
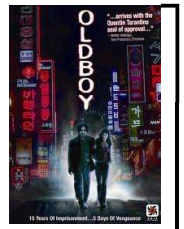
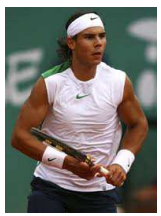
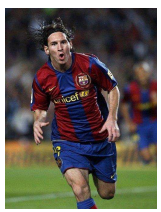
Set-up: Matrix $\Theta^* \in \mathbb{R}^{p_1 \times p_2}$ with rank $r \ll \min\{p_1, p_2\}$.

Estimator:

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

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 ...

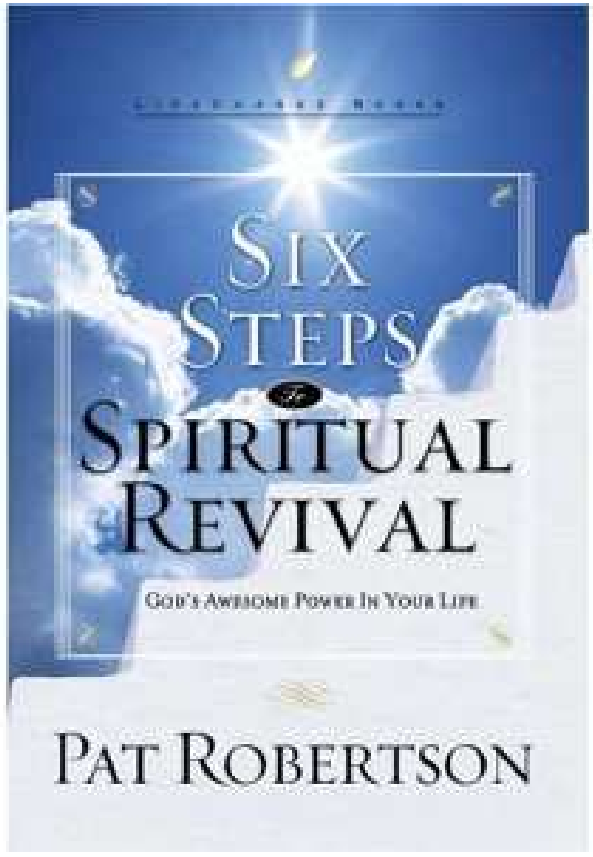
Application: Collaborative filtering

						
	4	$*$	3	$\ldots$	$\ldots$	$*$
	3	5	$*$	$\ldots$	$\ldots$	2
	5	4	3	$\ldots$	$\ldots$	3
	2	$*$	$*$	$\ldots$	$\ldots$	1

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

(e.g., Srebro, Alon & Jaakkola, 2004)

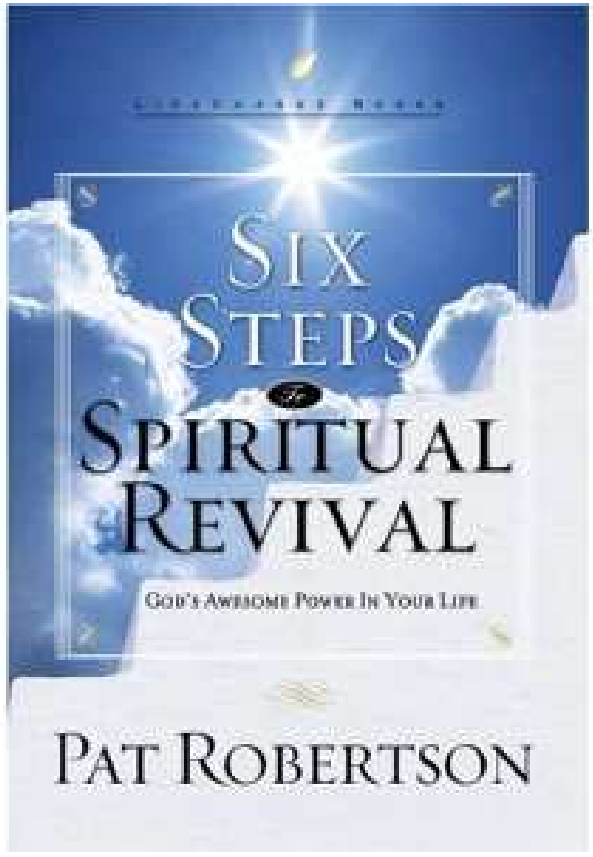
Security and robustness issues



Spiritual guide

Break-down of Amazon recommendation system, 2002.

Security and robustness issues



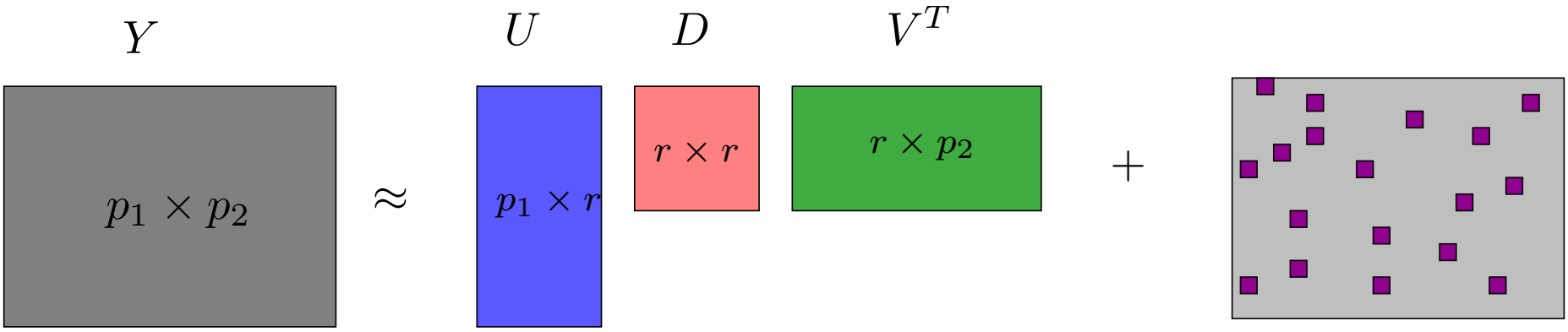
Spiritual guide

Sex manual

Break-down of Amazon recommendation system, 2002.

Matrix decomposition: Low-rank plus sparse

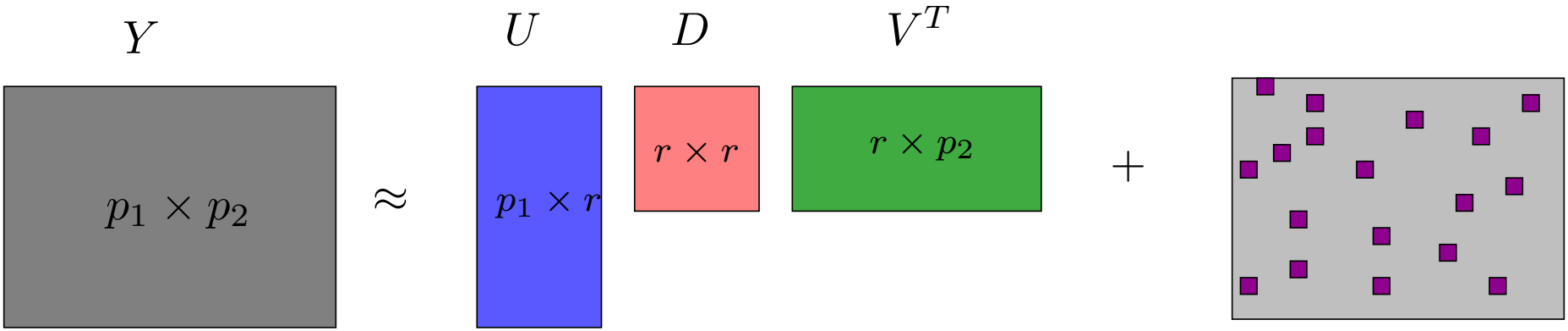
Matrix Y can be (approximately) decomposed into sum:



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

Matrix decomposition: Low-rank plus sparse

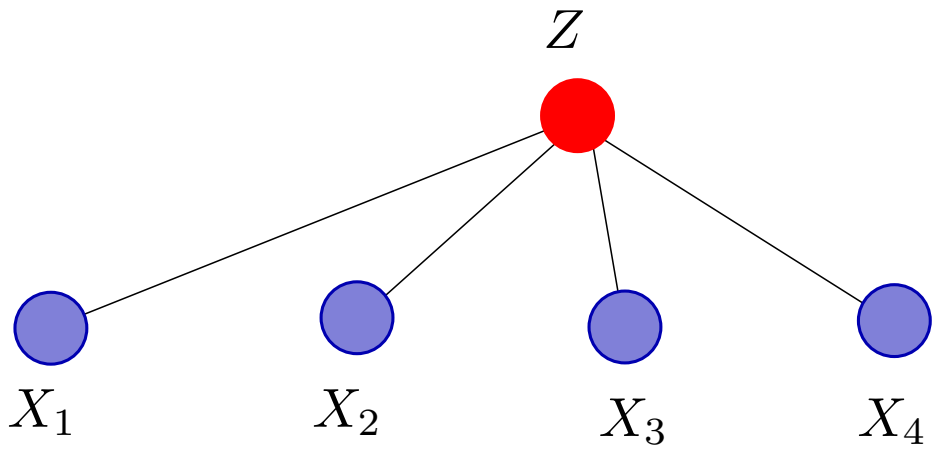
Matrix Y can be (approximately) decomposed into sum:



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

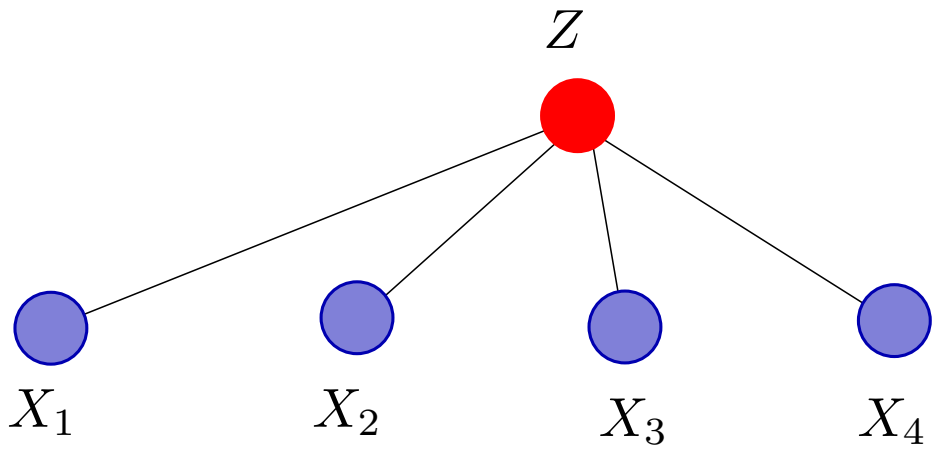
- exact decomposition: initially studied by Chandrasekaran, Sanghavi, Parillo & Willsky, 2009
- subsequent work: Candes et al., 2010; Xu et al., 2010 Hsu et al., 2010; Agarwal et al., 2011
- Various applications:
  - robust collaborative filtering
  - robust PCA
  - graphical model selection with hidden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

Gauss-Markov models with hidden variables



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

Gauss-Markov models with hidden variables



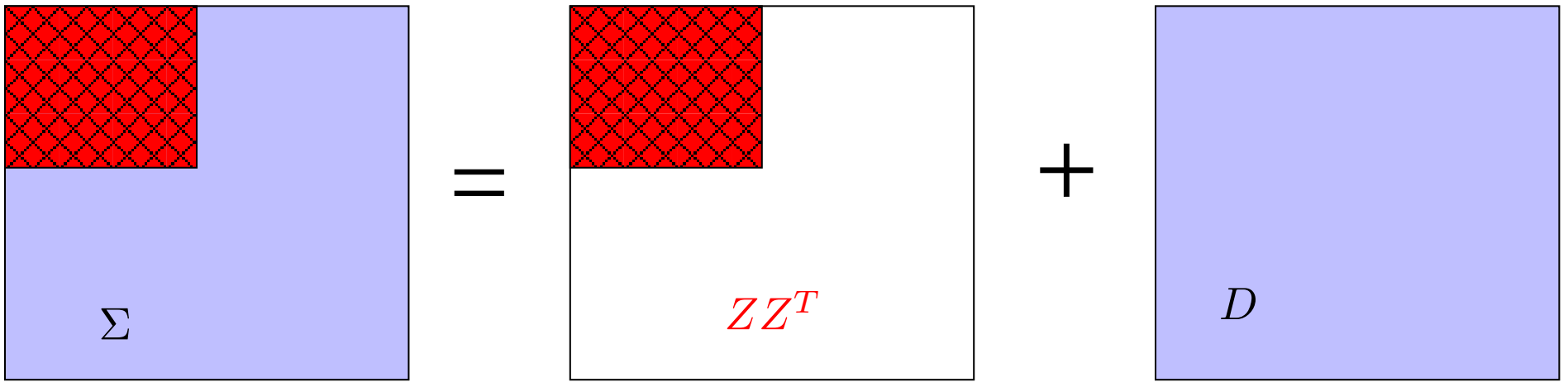
Problems with hidden variables: conditioned on hidden Z, vector $X = (X_1, X_2, X_3, X_4)$ is Gauss-Markov.

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

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

(Chandrasekaran, Parrilo & Willsky, 2010)

Example: Sparse principal components analysis



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

Estimator:

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

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 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?

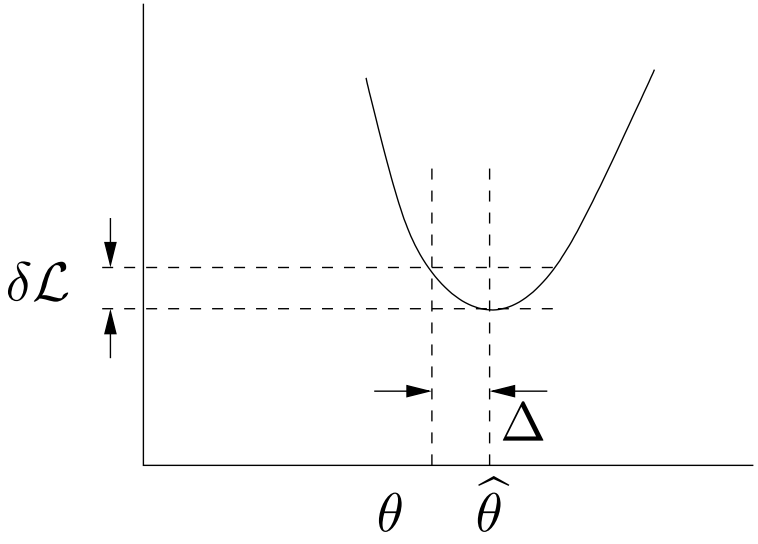
Answer: Yes, two essential ingredients.

(I) Restricted strong convexity of loss function

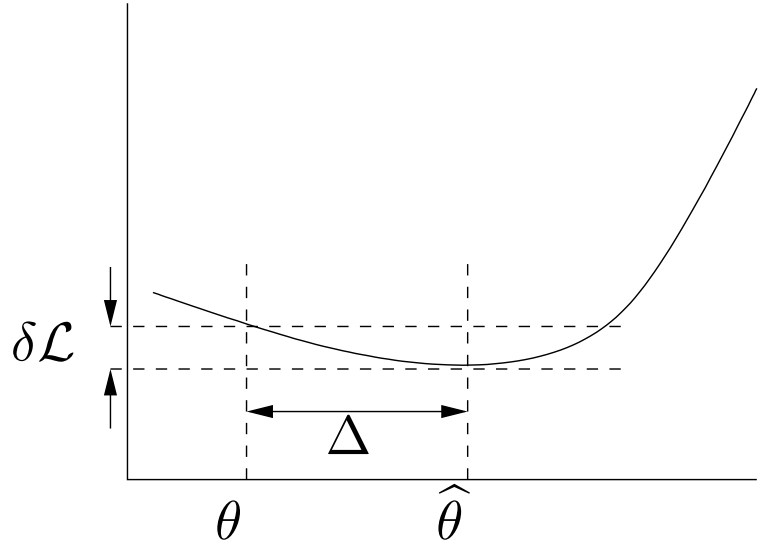
(II) Decomposability of the regularizer

(I) Role of curvature

1. Curvature controls difficulty of estimation:



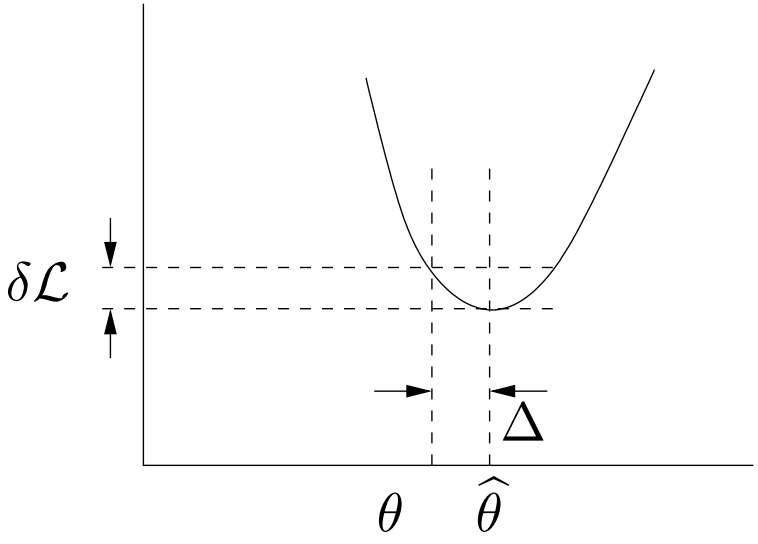
High curvature: easy to estimate



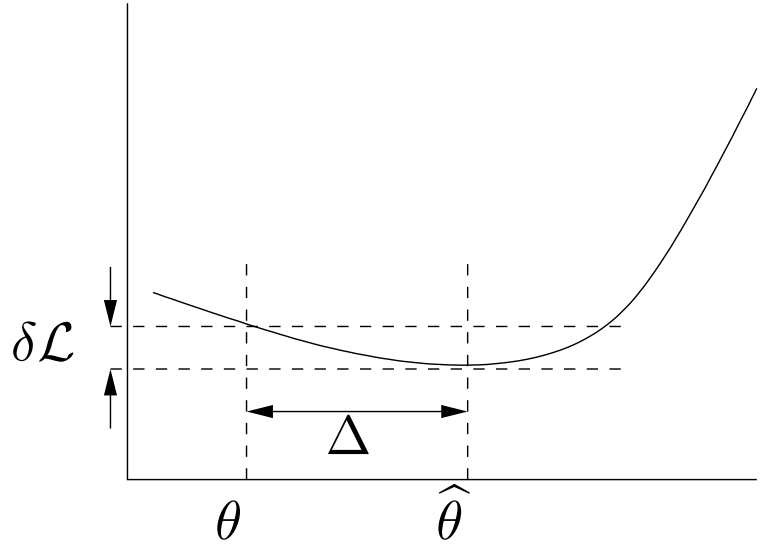
(b) Low curvature: harder

(I) Role of curvature

1. Curvature controls difficulty of estimation:



High curvature: easy to estimate



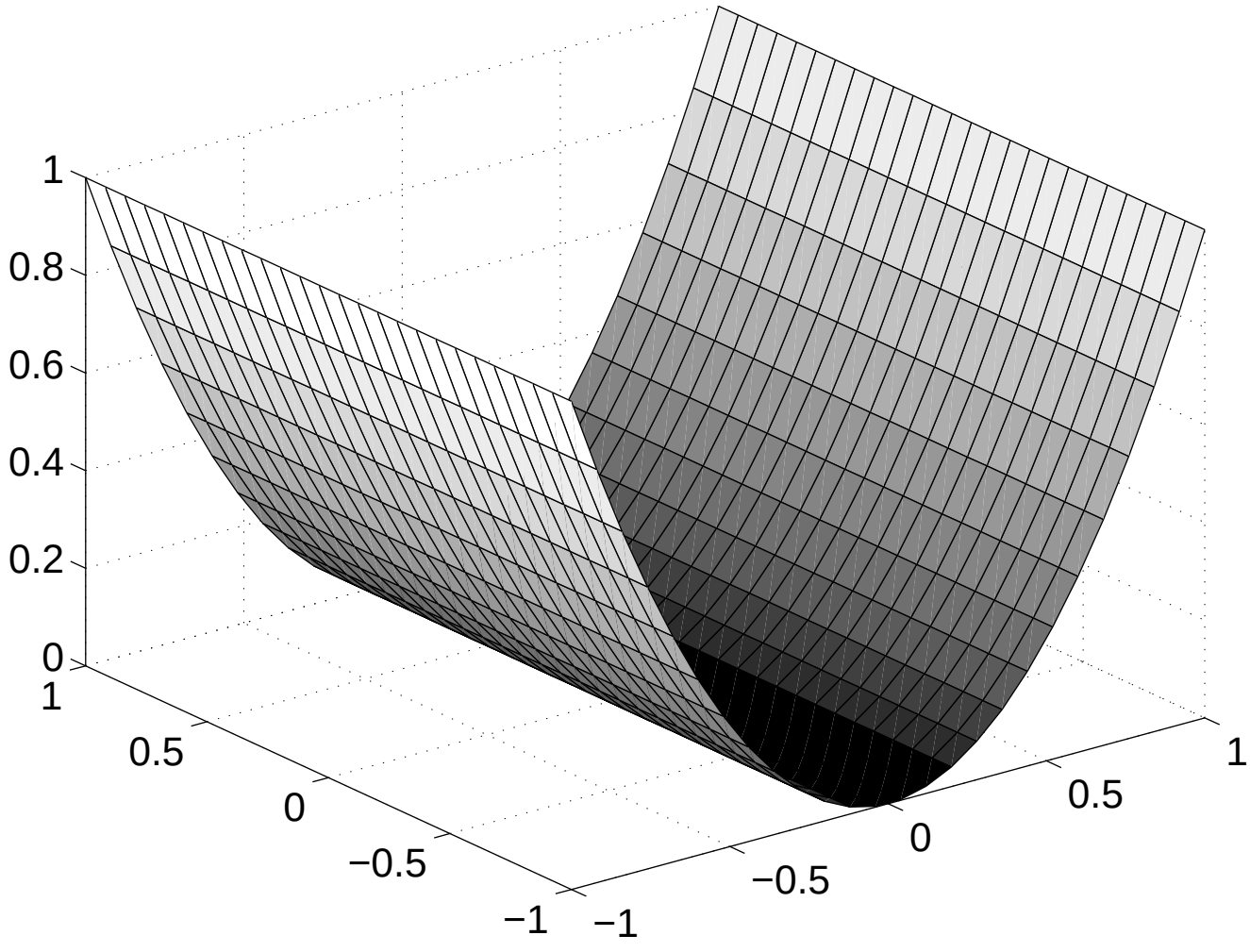
(b) Low curvature: harder

2. captured by lower bound on Taylor series error $\mathcal{T}_{\mathcal{L}}(\Delta; \theta^*)$

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

for all Δ around θ^*.

High dimensions: no strong convexity!



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

Restricted strong convexity

Definition

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

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

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

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

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 $p > n$

Restricted strong convexity

Definition

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

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

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 $p > n$

- RSC enforces a lower bound on curvature, but only when $\mathcal{R}^2(\Delta) \ll \|\Delta\|_e^2$

Least-squares: RSC $\equiv$ restricted eigenvalue

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

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

Least-squares: RSC $\equiv$ restricted eigenvalue

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

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

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

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

Least-squares: RSC $\equiv$ restricted eigenvalue

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

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

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

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

Least-squares: RSC $\equiv$ restricted eigenvalue

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

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

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

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

- holds with high probability for various sub-Gaussian designs when $n \gtrsim s \log p/s$.

- fairly strong dependency between covariates is possible

Restricted strong convexity for GLMS

- generalized linear model linking covariates $x \in \mathbb{R}^p$ to output $y \in \mathbb{R}$:

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

Restricted strong convexity for GLMS

- generalized linear model linking covariates $x \in \mathbb{R}^p$ to output $y \in \mathbb{R}$:

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

- Taylor series expansion involves random Hessian

$$H(\theta) = \frac{1}{n} \sum_{i=1}^{n} \Phi''\big(\langle \theta, X_i \rangle\big) X_i X_i^T \in \mathbb{R}^p.$$

Restricted strong convexity for GLMS

- generalized linear model linking covariates $x \in \mathbb{R}^p$ to output $y \in \mathbb{R}$:

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

- Taylor series expansion involves random Hessian

$$H(\theta) = \frac{1}{n} \sum_{i=1}^{n} \Phi''\big(\langle \theta, X_i \rangle\big) X_i X_i^T \in \mathbb{R}^p.$$

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

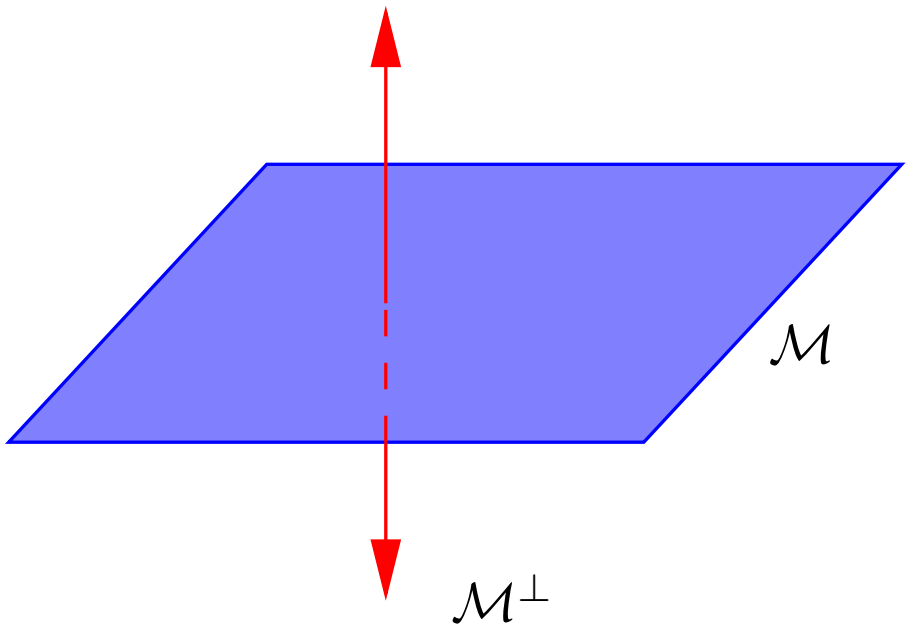
For zero-mean sub-Gaussian covariates X_i with covariance Σ and any GLM,

$$\mathcal{T}_{\mathcal{L}}(\Delta; \theta^*) \geq c_3 \|\Sigma^{\frac{1}{2}} \Delta\|_2 \left\{ \|\Sigma^{\frac{1}{2}} \Delta\|_2 - c_4 \kappa(\Sigma) \left(\frac{\log p}{n}\right)^{1/2} \|\Delta\|_1 \right\} \text{ for all } \|\Delta\|_2 \leq 1$$

with probability at least $1 - c_1 \exp(-c_2 n)$.

Here $\kappa(\Sigma) = \max_j \Sigma_{jj}$.

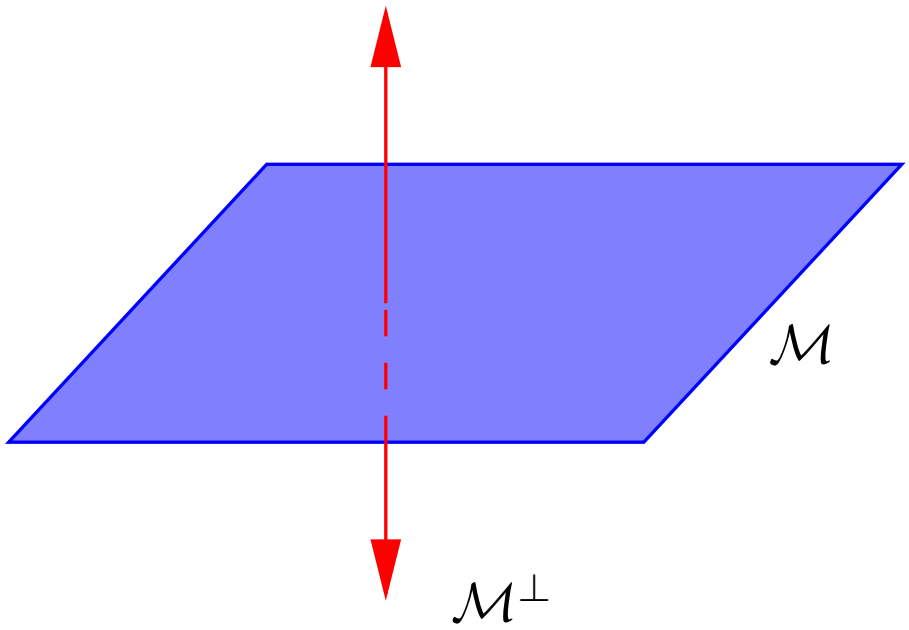
(II) Decomposable regularizers



Subspace $\mathcal{M}$: Approximation to model parameters

Complementary subspace $\mathcal{M}^{\perp}$: Undesirable deviations.

(II) Decomposable regularizers



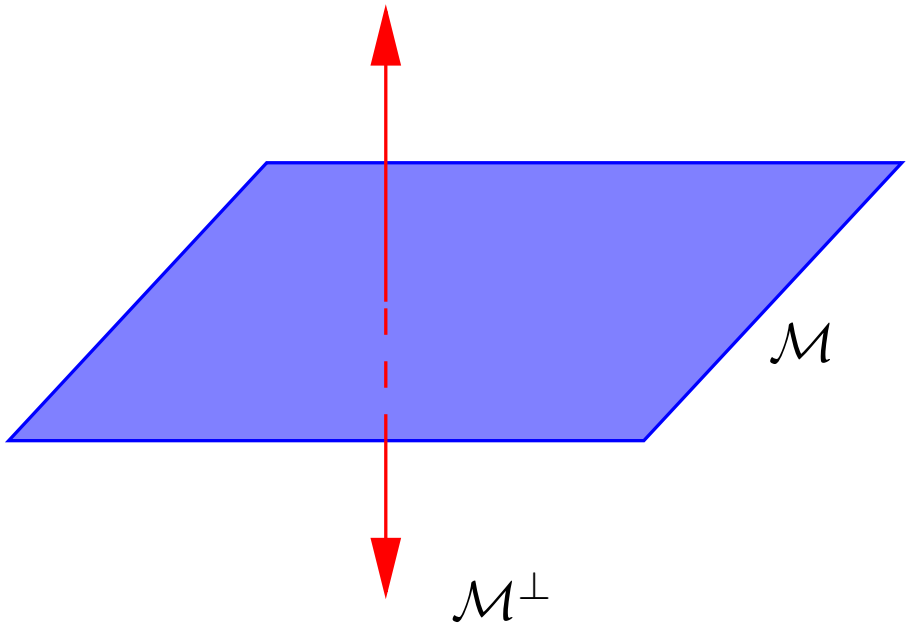
Subspace $\mathcal{M}$: Approximation to model parameters

Complementary subspace $\mathcal{M}^{\perp}$: Undesirable deviations.

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

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

(II) Decomposable regularizers



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

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

Includes:

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

Main theorem

Estimator

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

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

Main theorem

Estimator

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

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

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

Suppose that $\theta^* \in \mathcal{M}$. 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

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|^2 \precsim \frac{1}{\gamma^2(\mathcal{L})} \left\{ \lambda_n^2 + \tau^2(\mathcal{L}) \right\} \Psi^2(\mathcal{M}).$$

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.: $\Psi(\mathcal{M}) = \sup_{\theta \in \mathcal{M} \setminus \{0\}} \mathcal{R}(\theta)/\|\theta\|$

Main theorem

Estimator

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

Theorem (Oracle version)

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

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\|^2 \precsim \underbrace{\frac{(\lambda_n')^2}{\gamma^2(\mathcal{L})} \Psi^2(\mathcal{M})}_{\textit{Estimation error}} + \underbrace{\frac{\lambda_n'}{\gamma(\mathcal{L})} \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))}_{\textit{Approximation error}} \qquad \textit{where } \rho' = \max\{\rho, \tau$$

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.: $\Psi(\mathcal{M}) = \sup_{\theta \in \mathcal{M} \setminus \{0\}} \mathcal{R}(\theta)/\|\theta\|$

Example: Linear regression (exact sparsity)

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

Corollary

Suppose that true parameter θ^* is exactly s-sparse. Under RSC and with $\lambda_n \geq 2 \| \frac{X^T w}{n} \|_\infty$, then any Lasso solution satisfies $\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{4}{\gamma^2} s\, \lambda_n^2$.

Example: Linear regression (exact sparsity)

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

Corollary

Suppose that true parameter θ^* is exactly s-sparse. Under RSC and with $\lambda_n \geq 2\| \frac{X^T w}{n} \|_\infty$, then any Lasso solution satisfies $\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{4}{\gamma^2} s\, \lambda_n^2$.

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

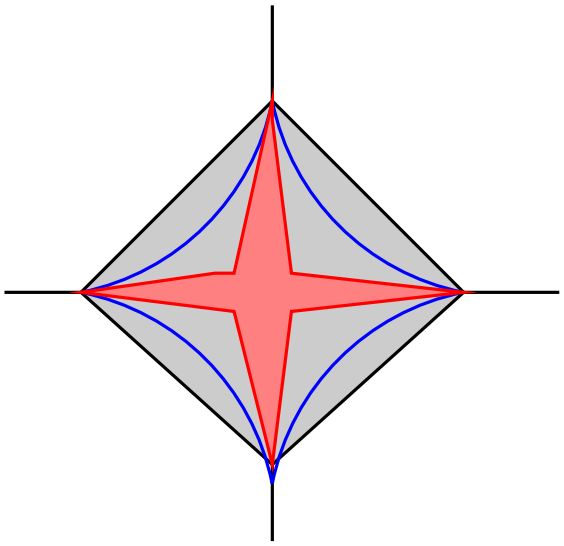
$$\| \frac{X^T w}{n} \|_\infty \leq \sqrt{\frac{3\sigma^2 \log p}{n}} \quad \text{w.h.p.} \implies \|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{12\sigma^2}{\gamma^2} \frac{s \log p}{n}.$$

(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"

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



Corollary

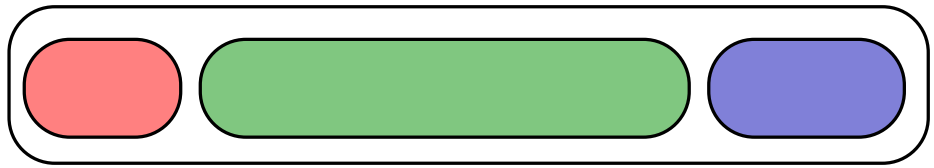
For $\theta^* \in \mathbb{B}_q(R_q)$, any Lasso solution satisfies (w.h.p.)

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

- rate known to be minimax optimal (Raskutti, W. & Yu, 2011)

Example: Group-structured regularizers

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



G_1 G_2 G_3

- 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; Bach et al., 2009; Lounici et al., 2009)

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*

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

Convergence rates for general group Lasso

Corollary

Say Θ^* *is supported on* $s_{\mathcal{G}}$ *groups, and* X *satisfies RSC. Then for regularization parameter*

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

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

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

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

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

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

Convergence rates for general group Lasso

Corollary

Say Θ^* is supported on $s_{\mathcal{G}}$ groups, and X satisfies RSC. Then for regularization parameter

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

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

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

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):

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

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):

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

- various applications:
  - matrix completion
  - rank-reduced multivariate regression
  - time-series modeling (vector autoregressions)
  - ...

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):

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

- various applications:
  - matrix completion
  - rank-reduced multivariate regression
  - time-series modeling (vector autoregressions)
  - ...

Some past work: Fazel, 2011; Srebro et al., 2004; Recht et al., 2007, Candes & Recht, 2008; Recht, 2009; Negahbah & W., 2010; Rohde & Tsybakov, 2010.

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., 2011)

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.*

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

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., 2011)

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.*

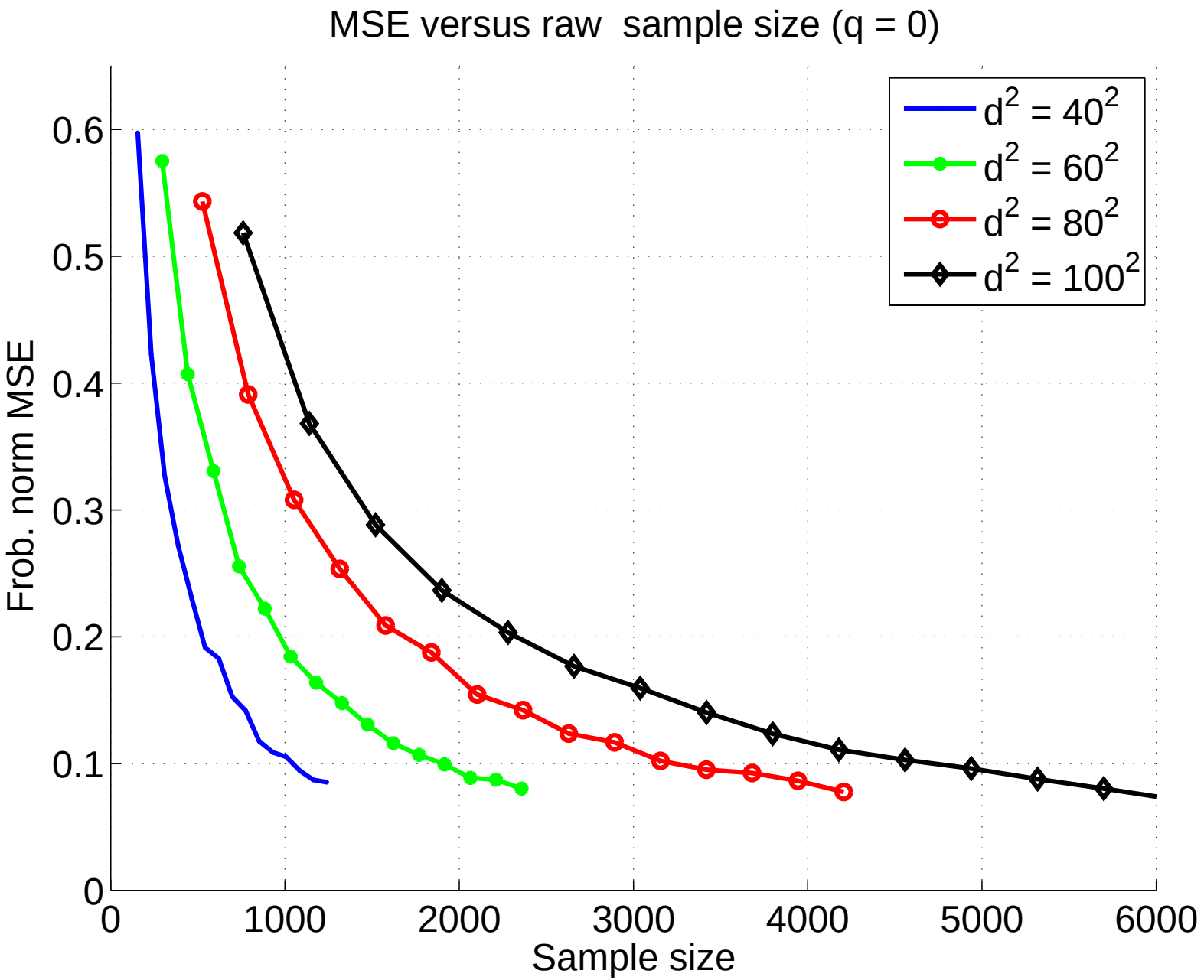
$$|||\widehat{\Theta} - \Theta^*|||_F^2 \quad \leq \quad c_0 \frac{R_q}{\gamma_\ell^2} \left(\frac{\sigma^2\,(p_1+p_2)}{n}\right)^{1-\frac{q}{2}}$$

- for a rank r matrix M

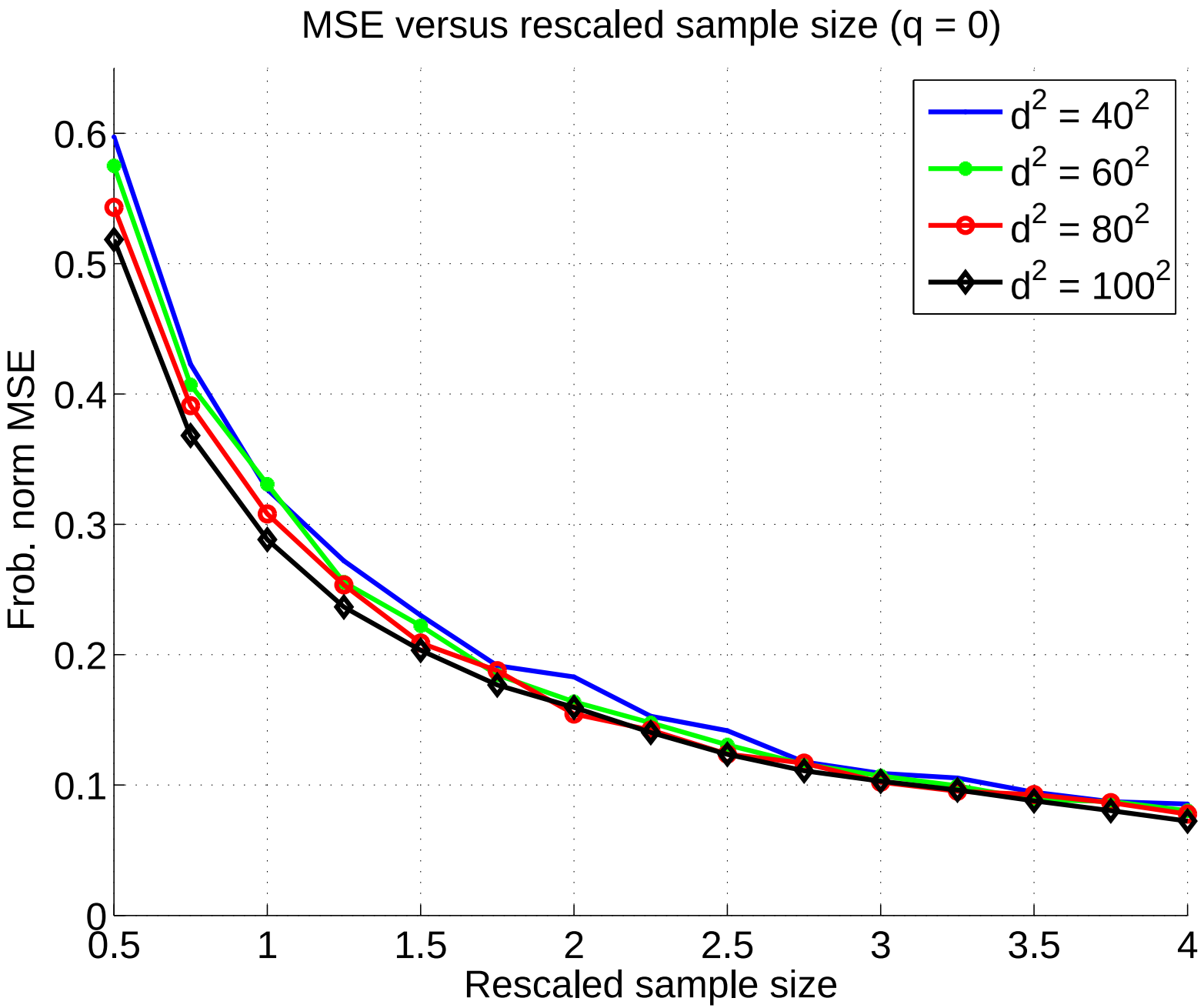
$$|||M|||_1 = \sum_{j=1}^{r} \sigma_j(M) \leq \sqrt{r}\sqrt{\sum_{j=1}^{r} \sigma_j^2(M)} \;=\; \sqrt{r}\,|||M|||_F$$

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

Noisy matrix completion (unrescaled)



Noisy matrix completion (rescaled)



Summary

- unified framework for high-dimensional M-estimators
  - decomposability of regularizer $\mathcal{R}$
  - restricted strong convexity of loss functions

- actual rates determined by:
  - noise measured in dual function $\mathcal{R}^*$
  - subspace constant Ψ in moving from $\mathcal{R}$ to error norm $\|\cdot\|$
  - restricted strong convexity constant

Summary

- unified framework for high-dimensional M-estimators
  - decomposability of regularizer $\mathcal{R}$
  - restricted strong convexity of loss functions

- actual rates determined by:
  - noise measured in dual function $\mathcal{R}^*$
  - subspace constant Ψ in moving from $\mathcal{R}$ to error norm $\|\cdot\|$
  - restricted strong convexity constant

Looking ahead to tomorrow:

From parametric to non-parametric problems: using kernel-based methods in high-dimensions.

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

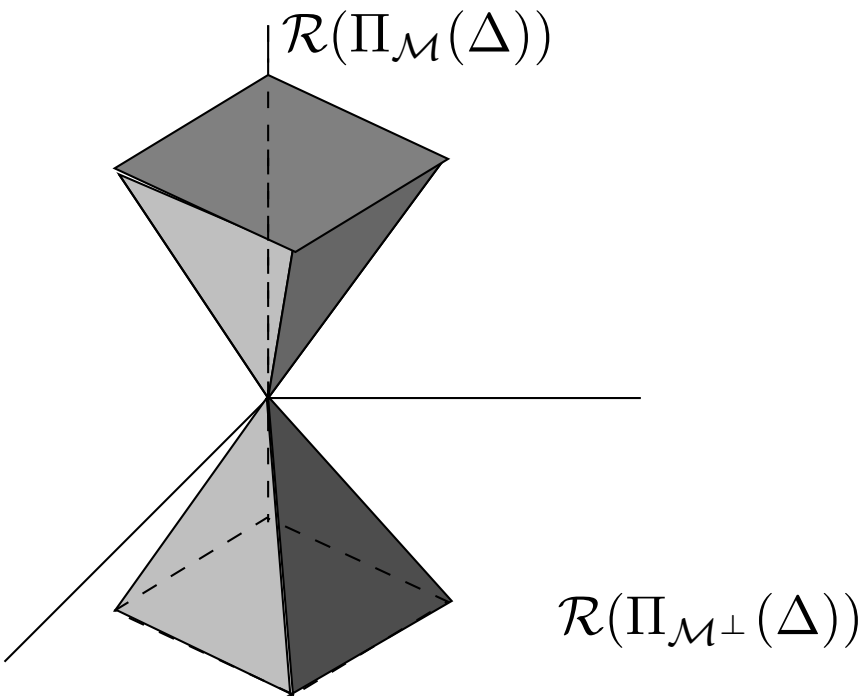
1. S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *Statistical Science*, December 2012.

2. S. Negahban and M. J. Wainwright (2011). Estimation rates of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(1):1069–1097.

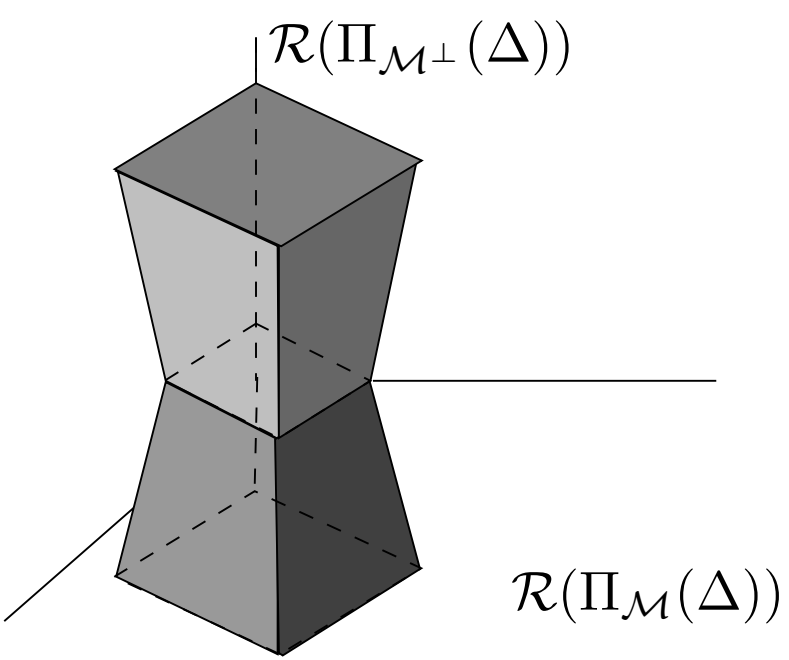
3. S. Negahban and M. J. Wainwright (2012). Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, May 2012.

4. G. Raskutti, M. J. Wainwright and B. Yu (2011) Minimax rates for linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10): 6976–6994.

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. $\mathcal{M}$. 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

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