A primer on high-dimensional statistics: Lecture 2

Martin Wainwright

UC Berkeley
Departments of Statistics, and EECS

Simons Institute Workshop, Bootcamp Tutorials

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:

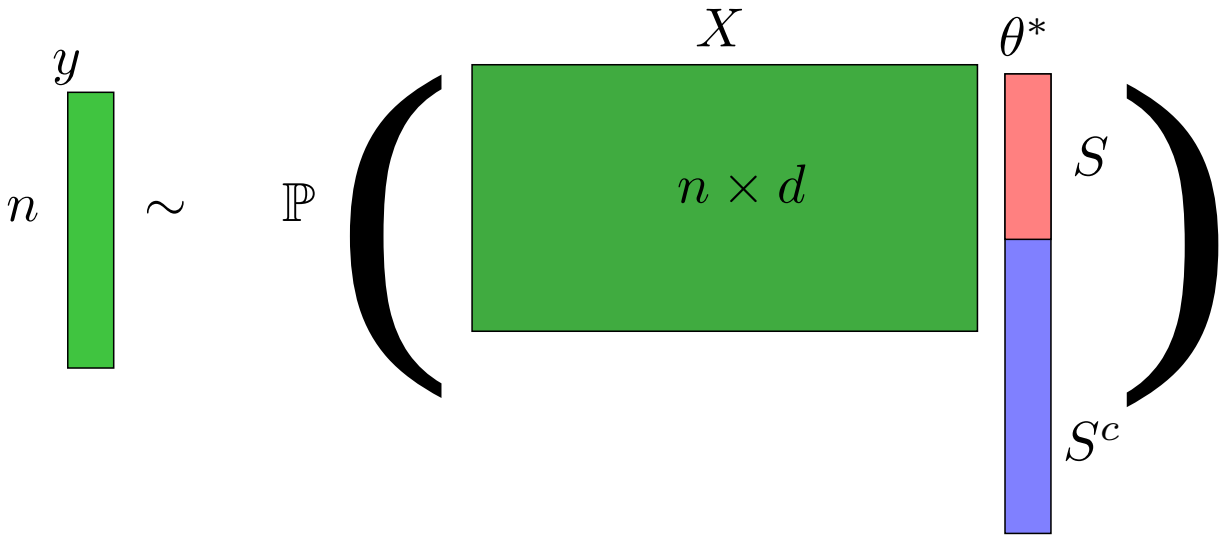
Is there a common set of underlying principles?

Last lecture (+): 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.

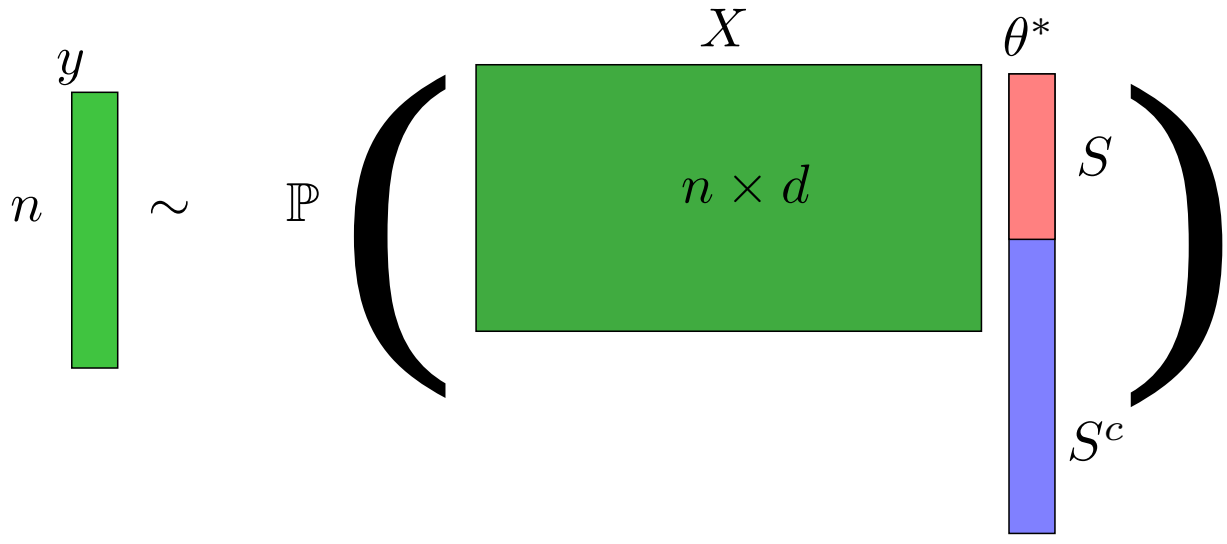


Last lecture (+): Sparse regression

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

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

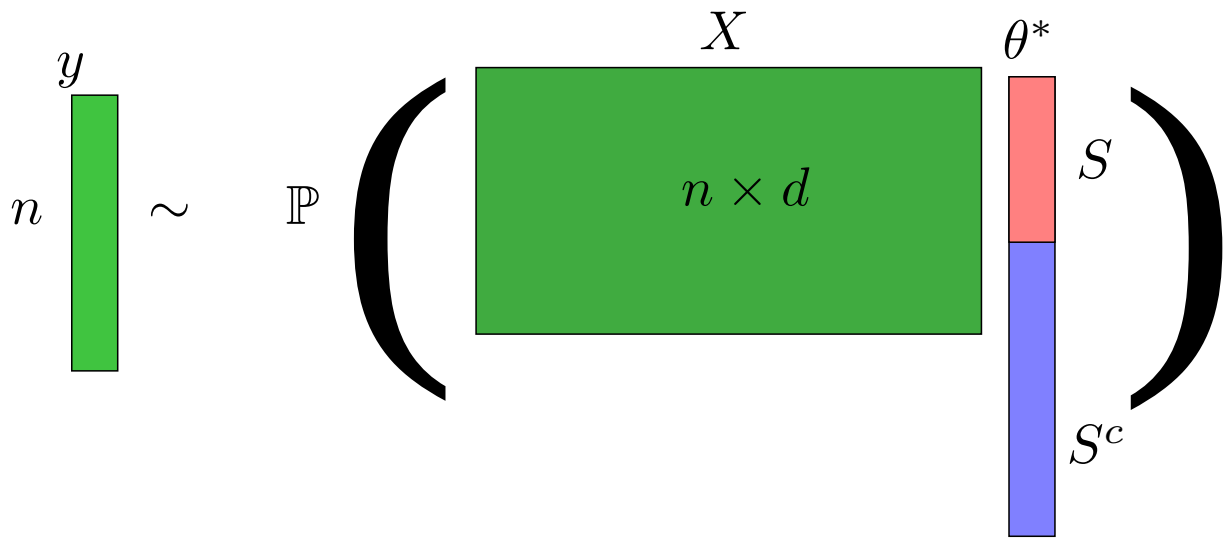
where $\theta \in \mathbb{R}^d$ is sparse.



Estimator: ℓ_1-regularized likelihood

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

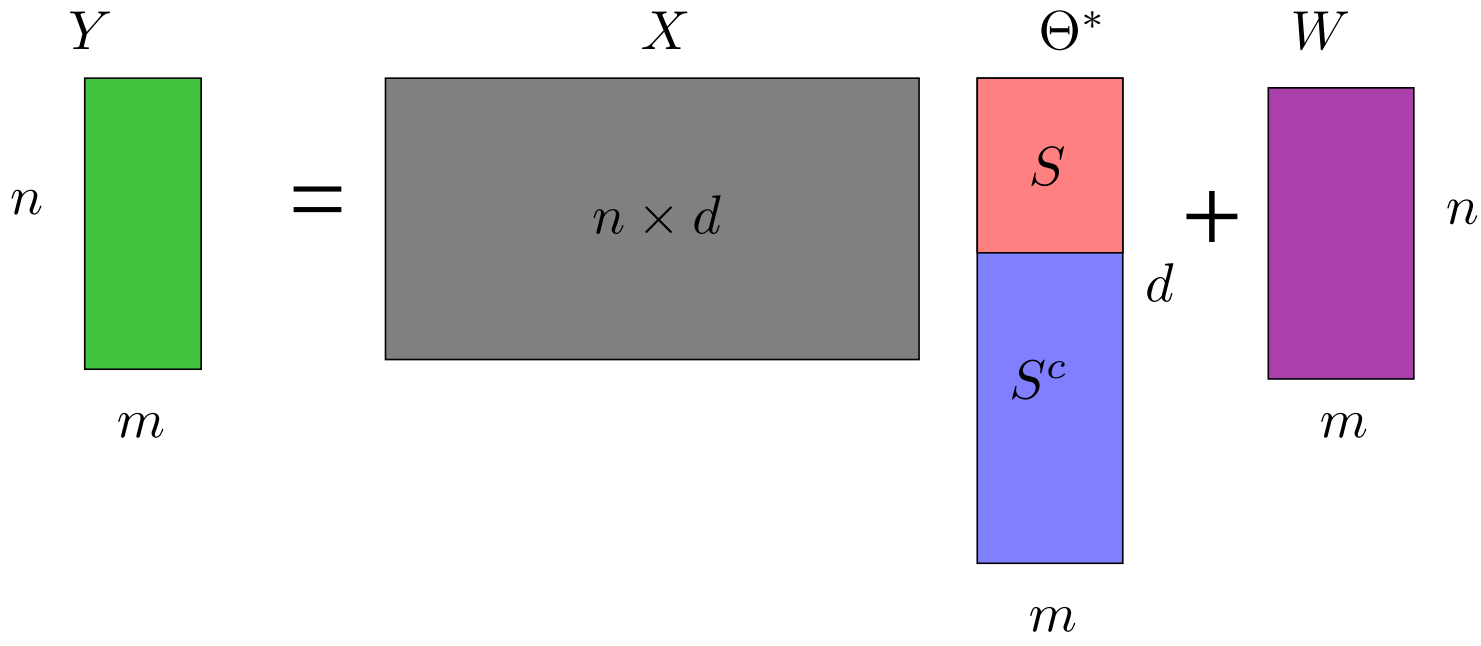
Last lecture (+): Sparse regression



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

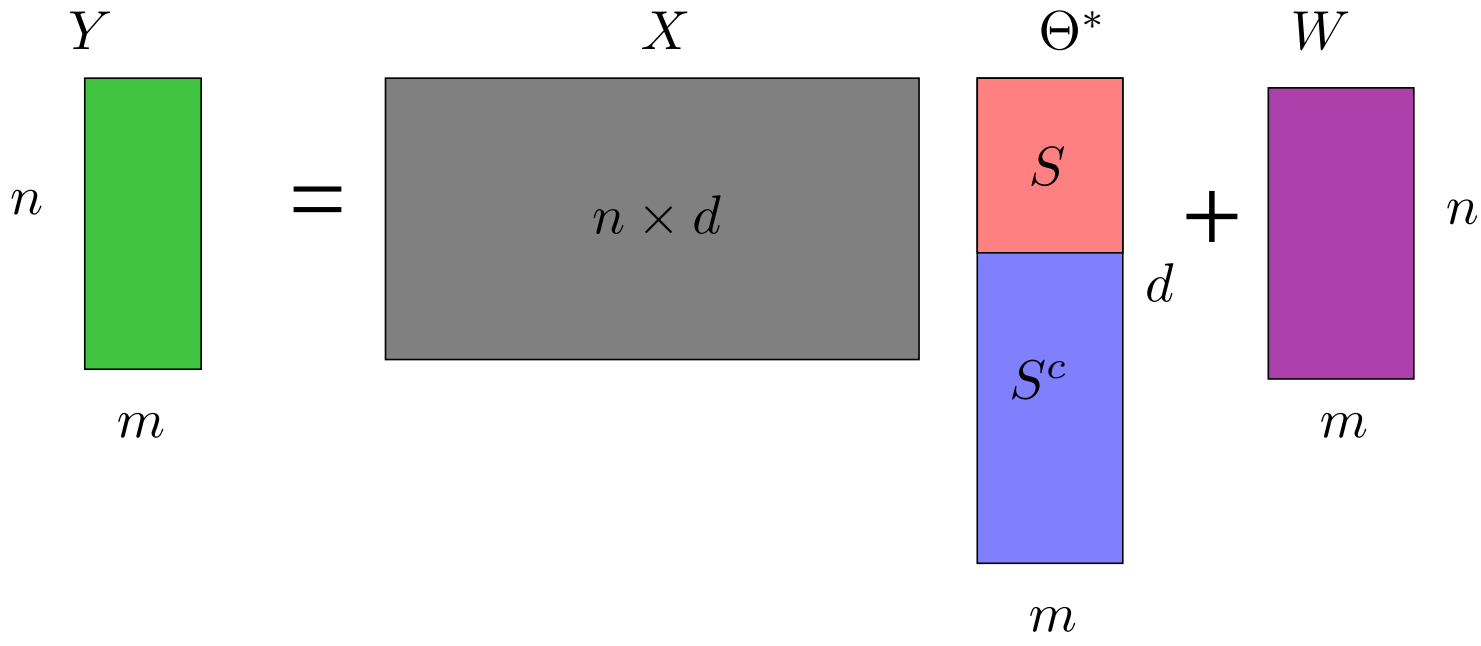
$$\widehat{\theta} \in \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ \log(1 + e^{\langle x_i, \theta \rangle}) - y_i \langle x_i, \theta \rangle \right\} + \lambda_n \|\theta\|_1 \right\}.$$

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.

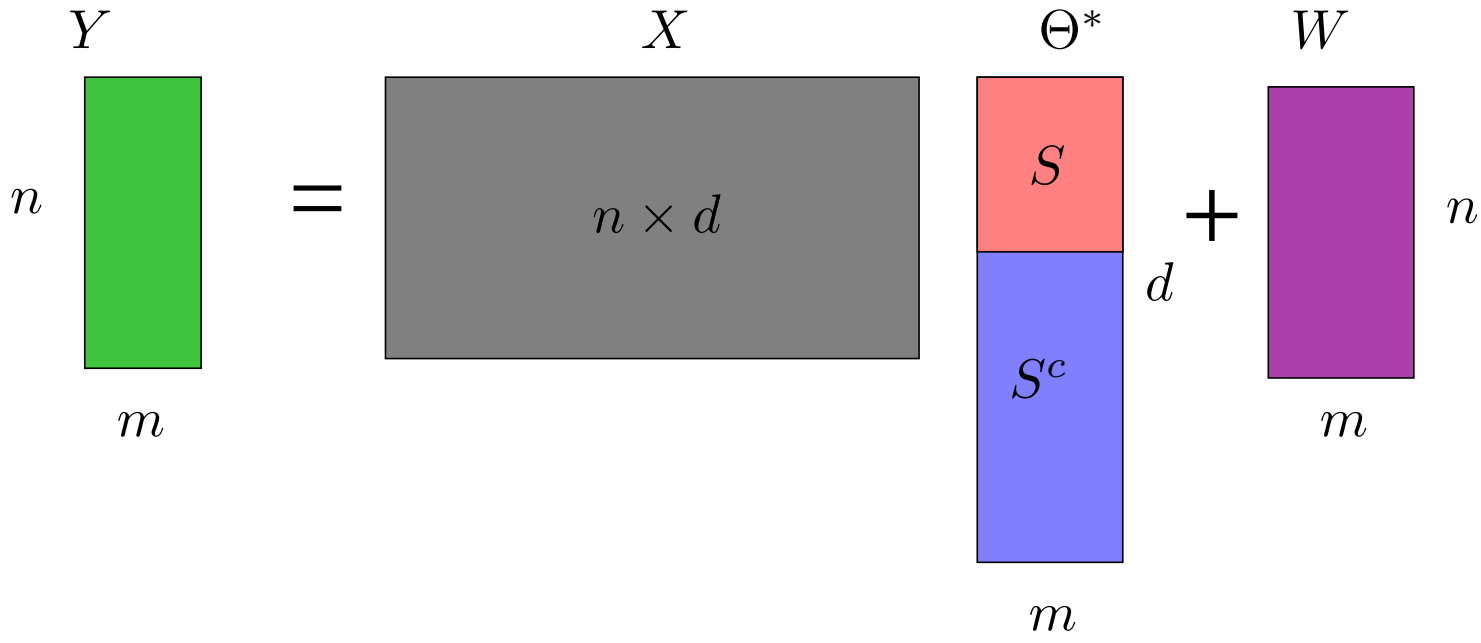
Example: Block sparsity and group Lasso



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

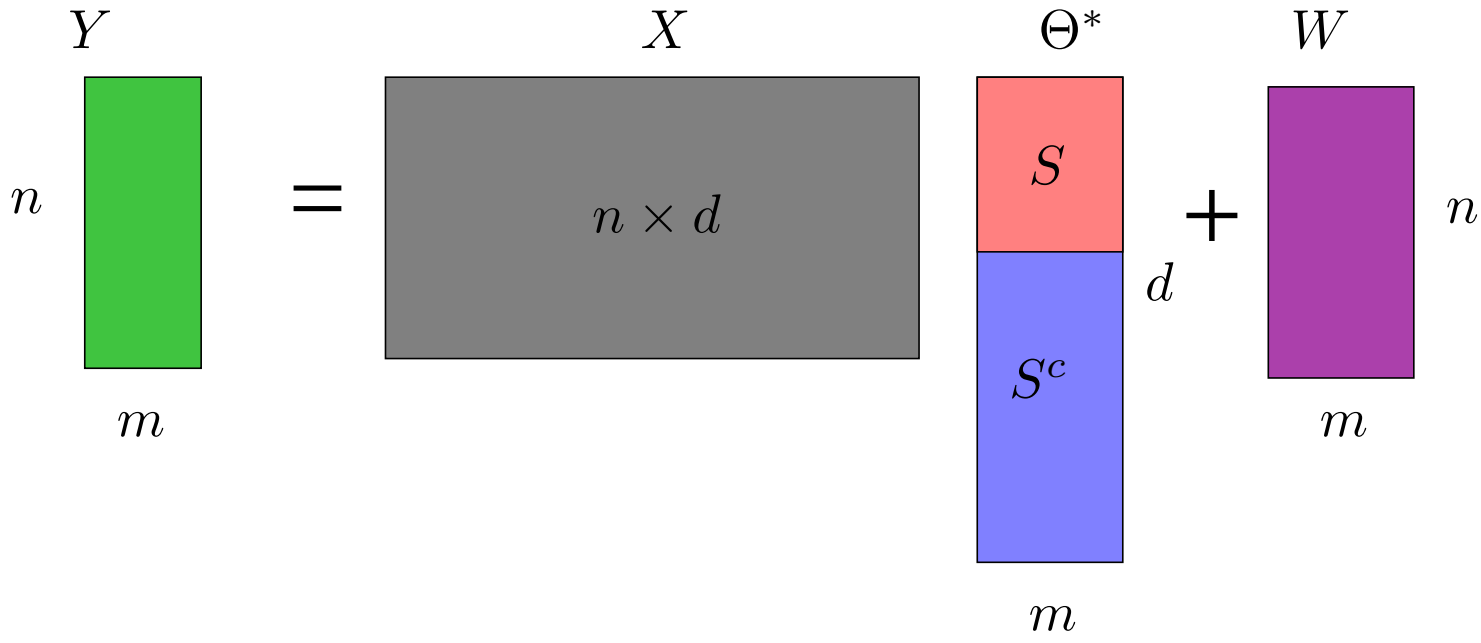
Example: Block sparsity and group Lasso



- 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

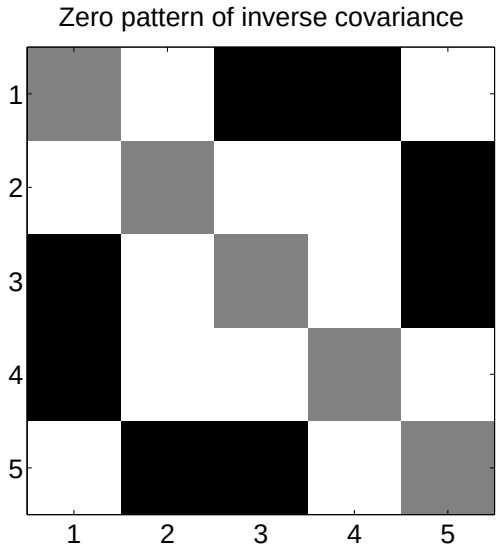


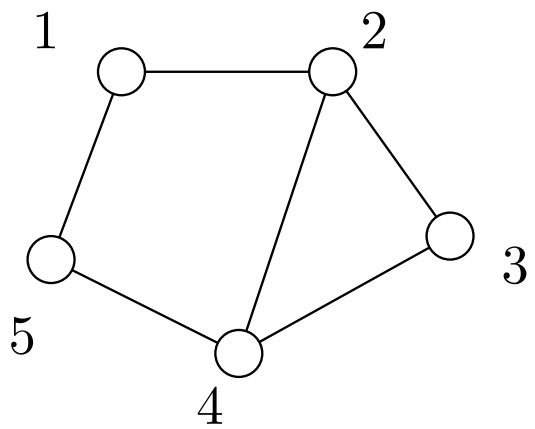
- Row-wise ℓ_1/ℓ_2-norm $|||\Theta|||_{1,2} = \sum_{j=1}^{d} \|\Theta_j\|_2$
- Weighted r-group Lasso: (Wright et al., 2005; Tropp et al., 2006; Yuan & Lin, 2006)

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

- 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





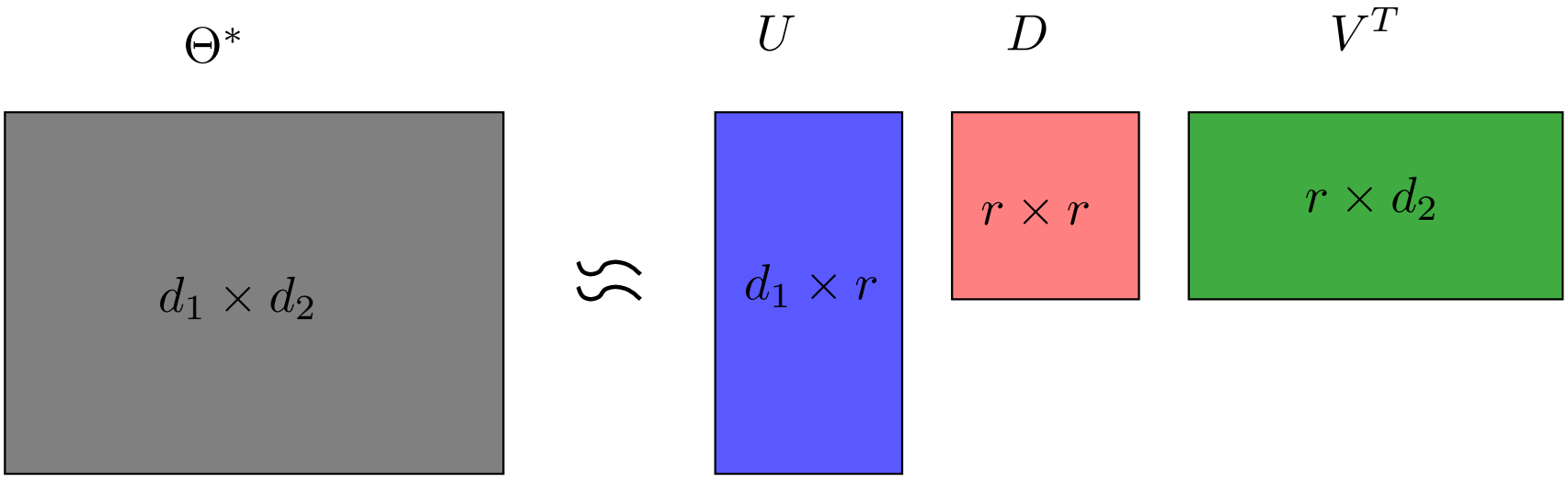
Set-up: Samples from random vector with sparse covariance Σ or sparse inverse covariance $\Theta^* \in \mathbb{R}^{d \times d}$.

Estimator (for inverse covariance)

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

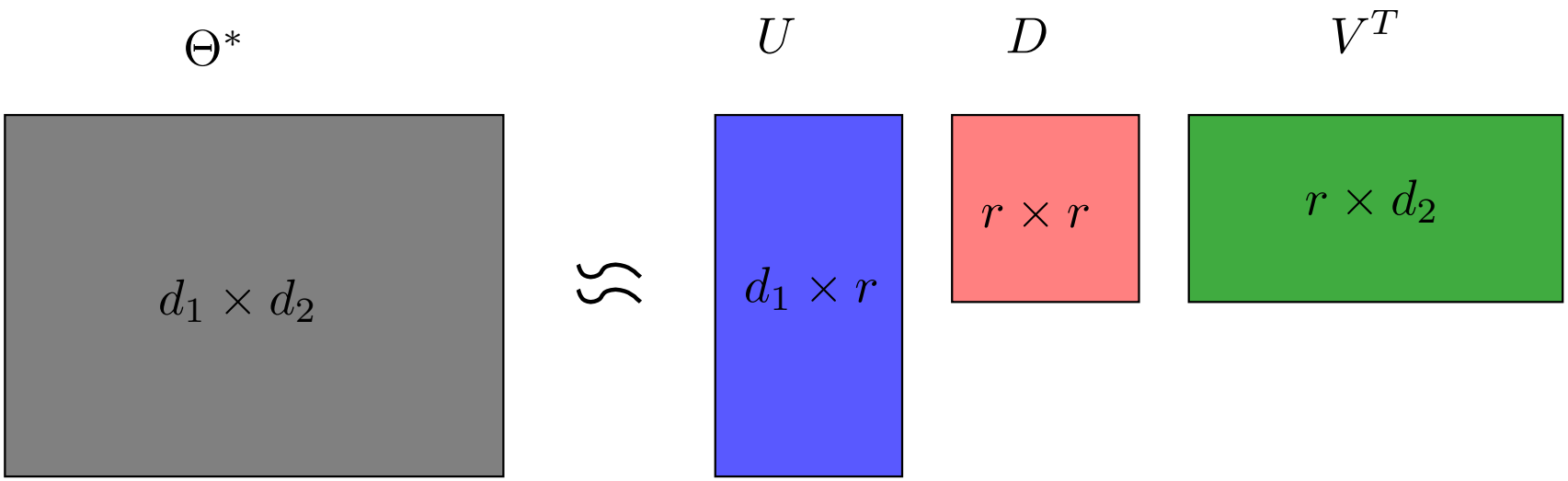
Some past work: Yuan & Lin, 2006; d'Asprémont et al., 2007; Bickel & Levina, 2007; El Karoui, 2007; d'Aspremont et al., 2007; Rothman et al., 2007; Zhou et al., 2007; Friedman et al., 2008; Lam & Fan, 2008; Ravikumar et al., 2008; Zhou, Cai & Huang, 2009; Guo et

Example: Low-rank matrix approximation



Set-up: Matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ with rank $r \ll \min\{d_1, d_2\}$.

Example: Low-rank matrix approximation



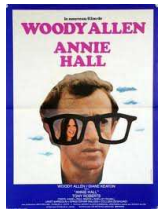
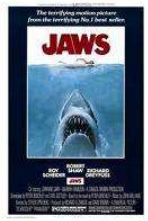
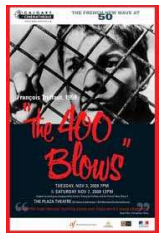
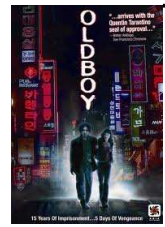
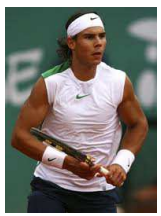
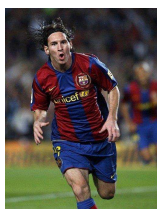
Set-up: Matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ with rank $r \ll \min\{d_1, d_2\}$.

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

$$\widehat{\Theta} \in \arg\min_{\Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\langle X_i, \Theta \rangle\rangle)^2 + \lambda_n \sum_{j=1}^{\min\{d_1, d_2\}} \gamma_j(\Theta) \right\}$$

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, Koltchinski et al., 2011

Application: Collaborative filtering

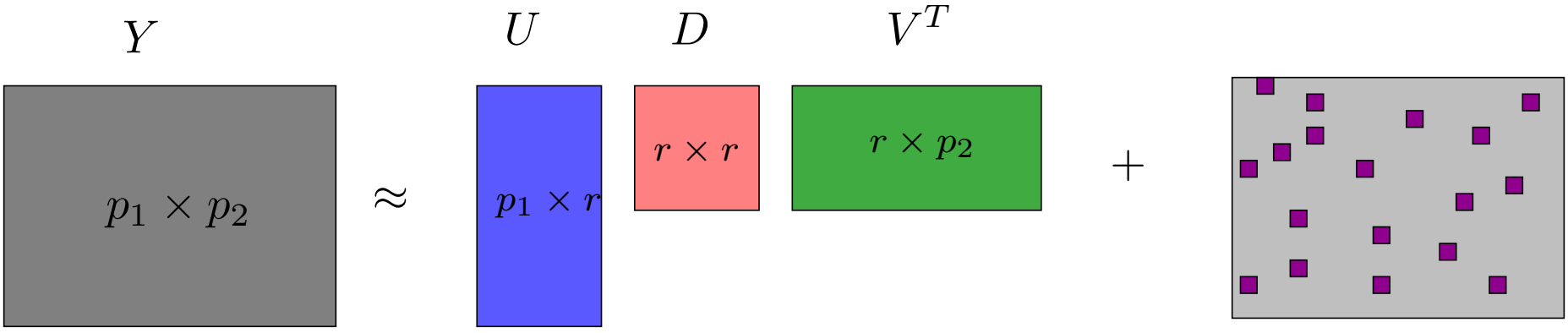
						
	4	$*$	3	$\ldots$	$\ldots$	$*$
	3	5	$*$	$\ldots$	$\ldots$	2
	5	4	3	$\ldots$	$\ldots$	3
	2	$*$	$*$	$\ldots$	$\ldots$	1

Universe of d_1 individuals and d_2 films Observe $n \ll d_2 d_2$ ratings

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

Example: Additive matrix decomposition

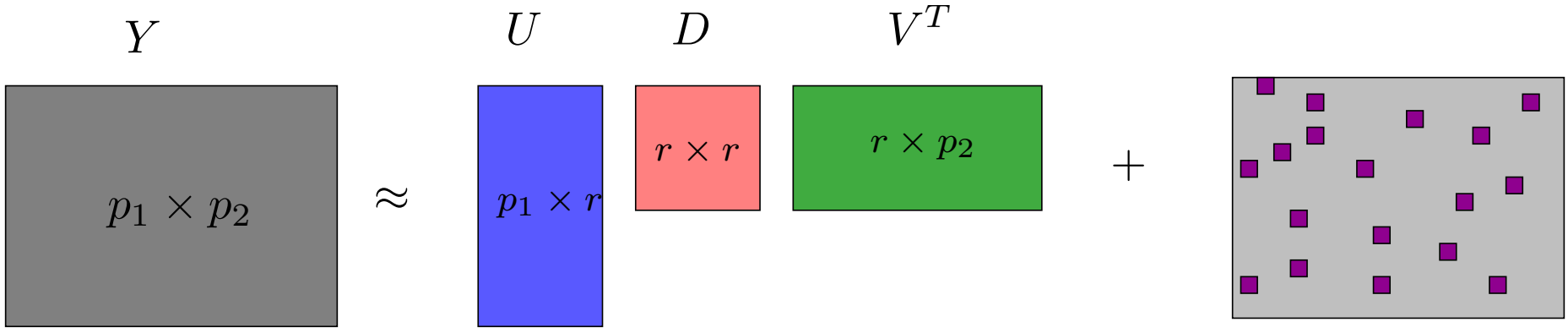
Matrix Y can be (approximately) decomposed into sum:



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

Example: Additive matrix decomposition

Matrix Y can be (approximately) decomposed into sum:

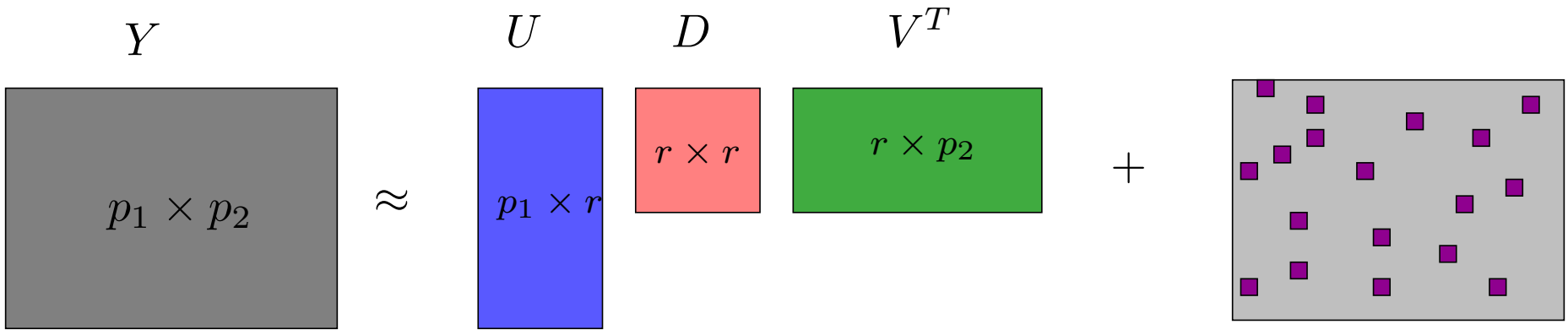


$$Y = \underbrace{\Theta^*}_{\text{Low-rank component}} + \underbrace{\Gamma^*}_{\text{Sparse component}}$$

- 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

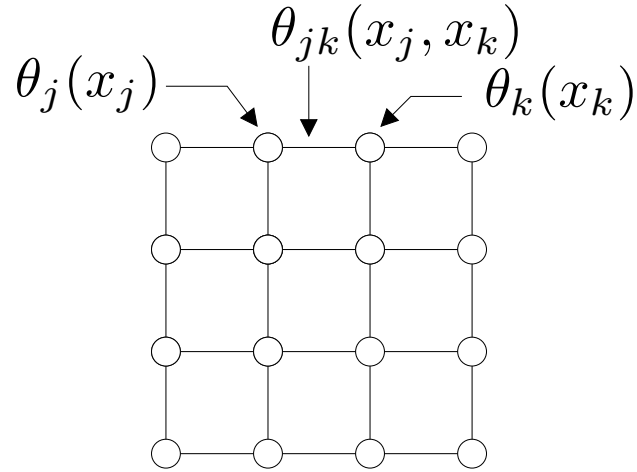
Matrix Y can be (approximately) decomposed into sum:



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

- 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:

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

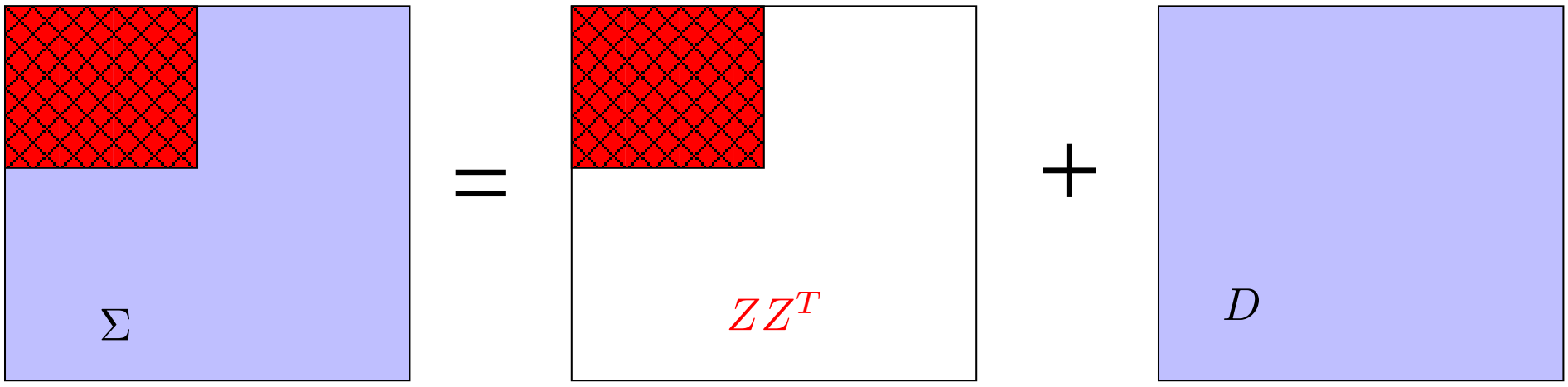
(Lin & Zhang, 2003; Meier et al., 2007; Ravikumar et al. 2007; Koltchinski and Yuan, 2008; Raskutti, W., & Yu, 2010)

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

Estimator:

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

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; 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?

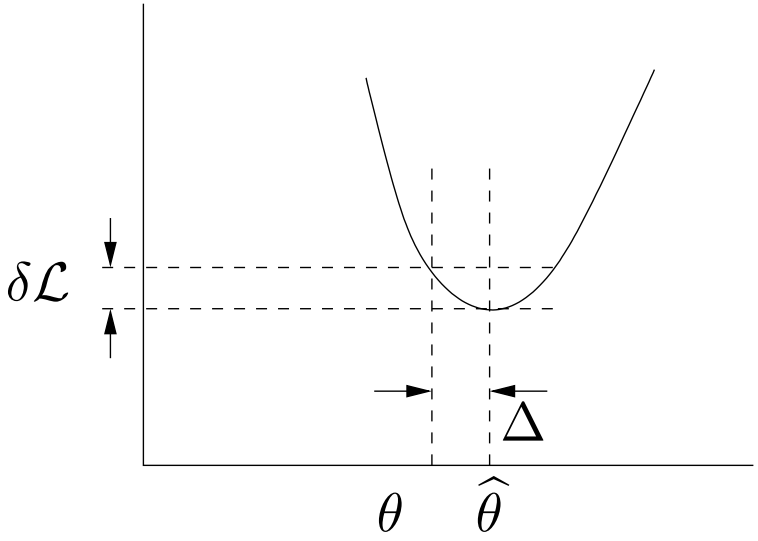
Answer: Yes, two essential ingredients.

(I) Restricted strong convexity of loss function

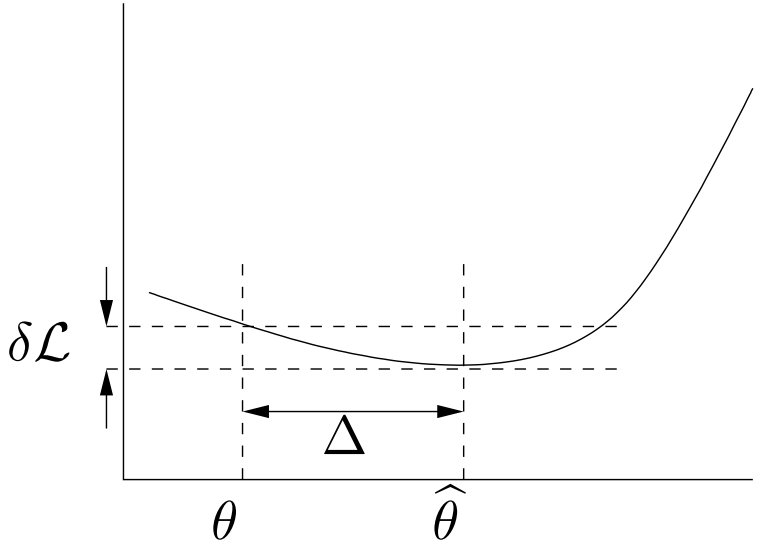
(II) Decomposability of the regularizer

(I) Classical role of curvature in statistics

1. Curvature controls difficulty of estimation:



High curvature: easy to estimate



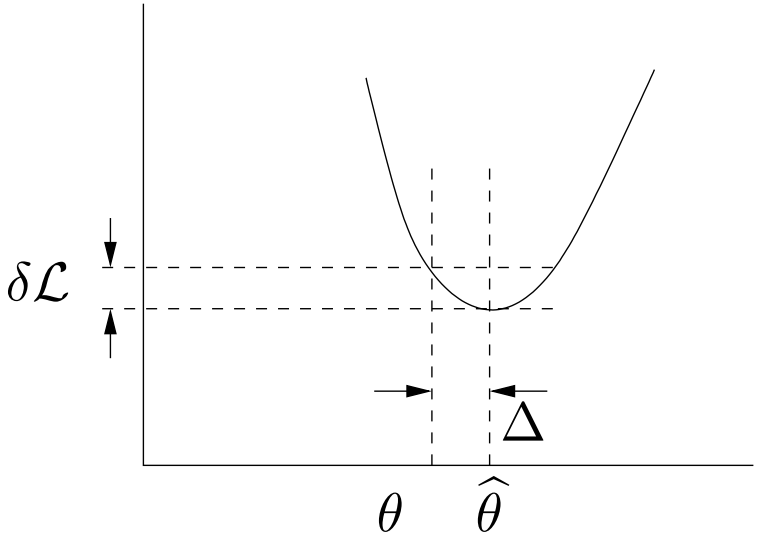
(b) Low curvature: harder

Canonical example:

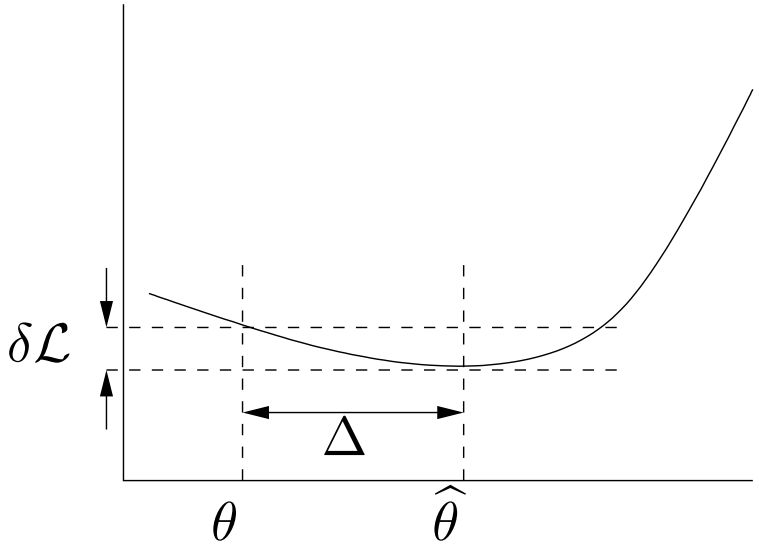
Log likelihood, Fisher information matrix and Cramér-Rao bound.

(I) Classical role of curvature in statistics

1. Curvature controls difficulty of estimation:



High curvature: easy to estimate



(b) Low curvature: harder

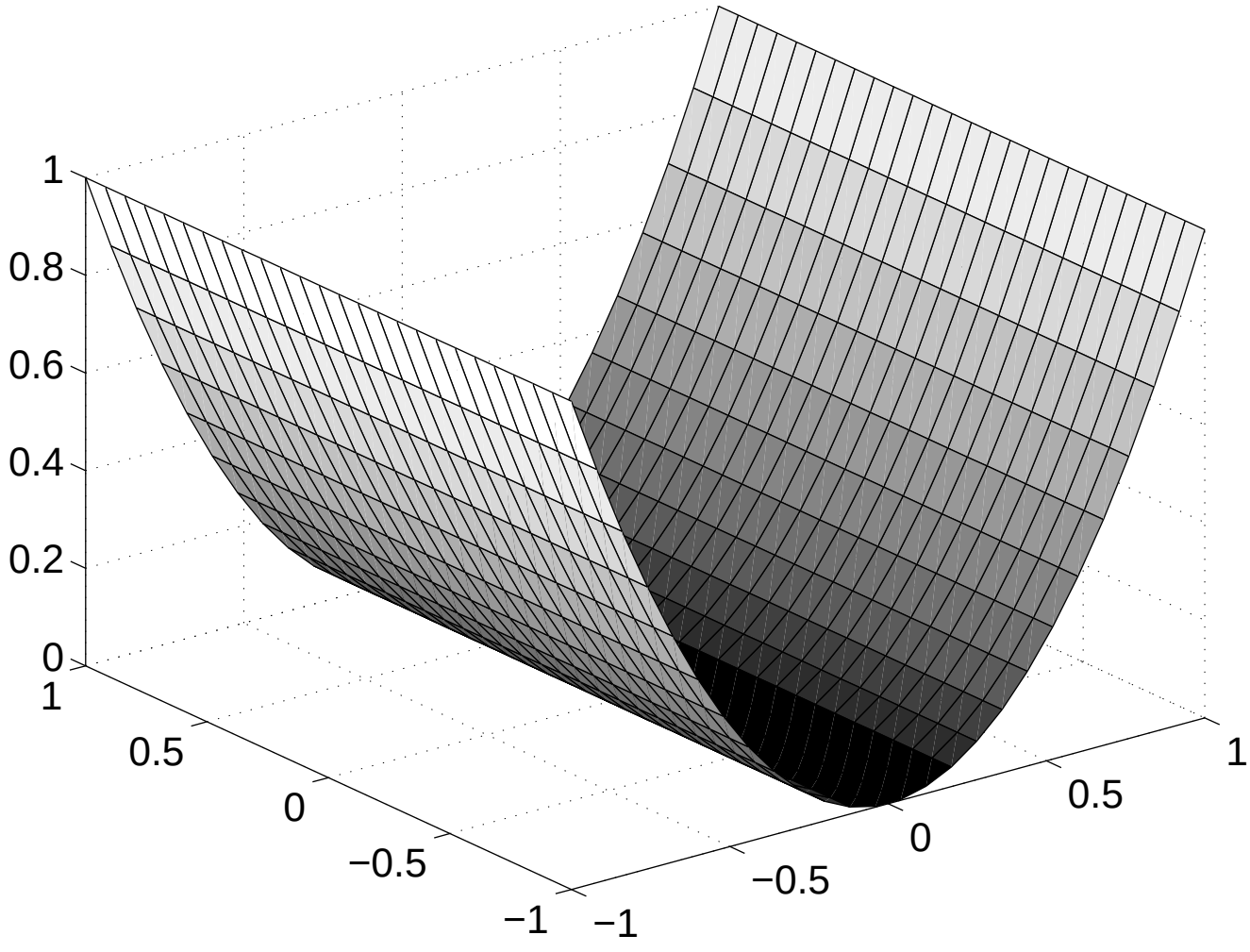
Canonical example:

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

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

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

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

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

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

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

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 $d > n$

Restricted strong convexity

Definition

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

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

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

Restricted strong convexity

Definition

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

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

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

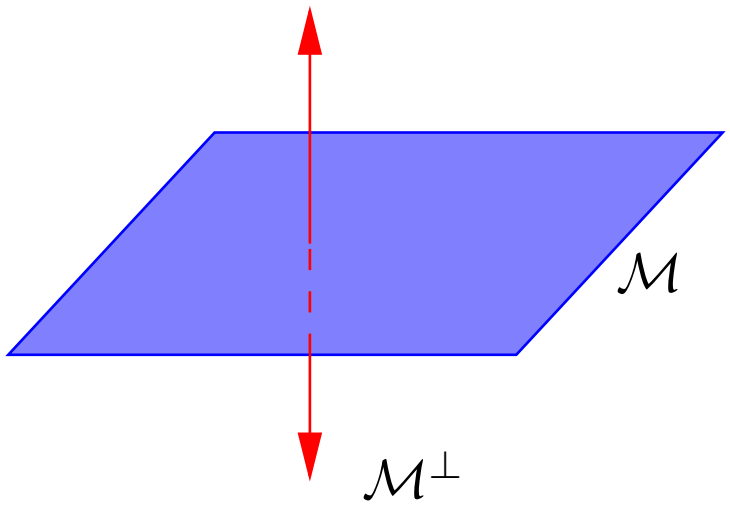
Theorem (Negahban et al., 2010; W., 2012)

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

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

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

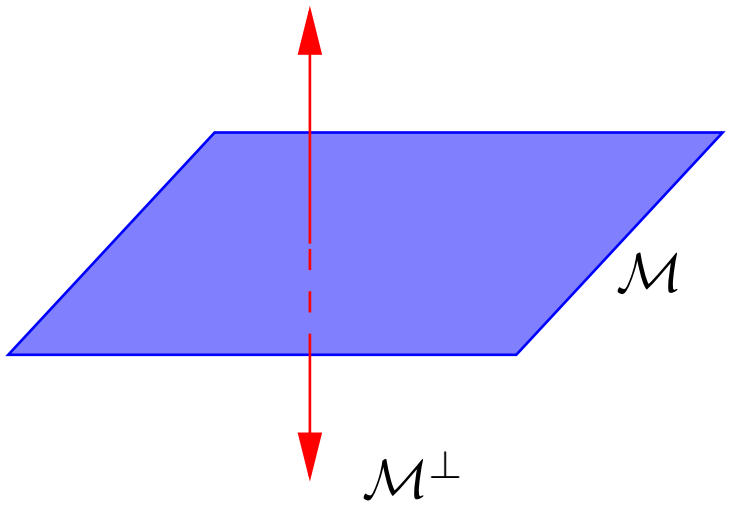
(II) Decomposable regularizers



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

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

(II) Decomposable regularizers



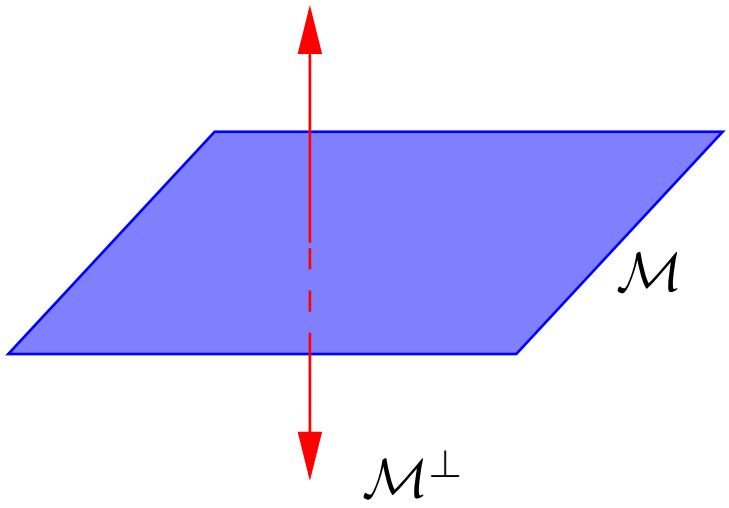
Subspace $\mathcal{M}$: Approximation to model parameters

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



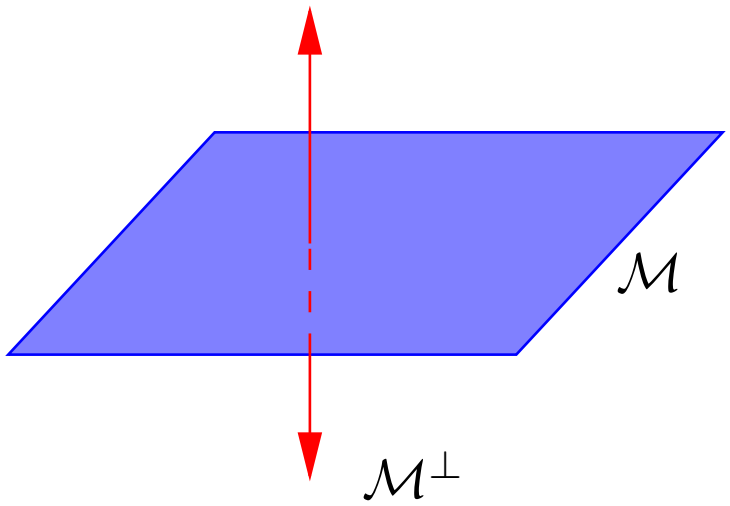
Regularizer $\mathcal{R}$ decomposes across $(\mathcal{M}, \mathcal{M}^{\perp})$ if

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

Includes:

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

(II) Decomposable regularizers



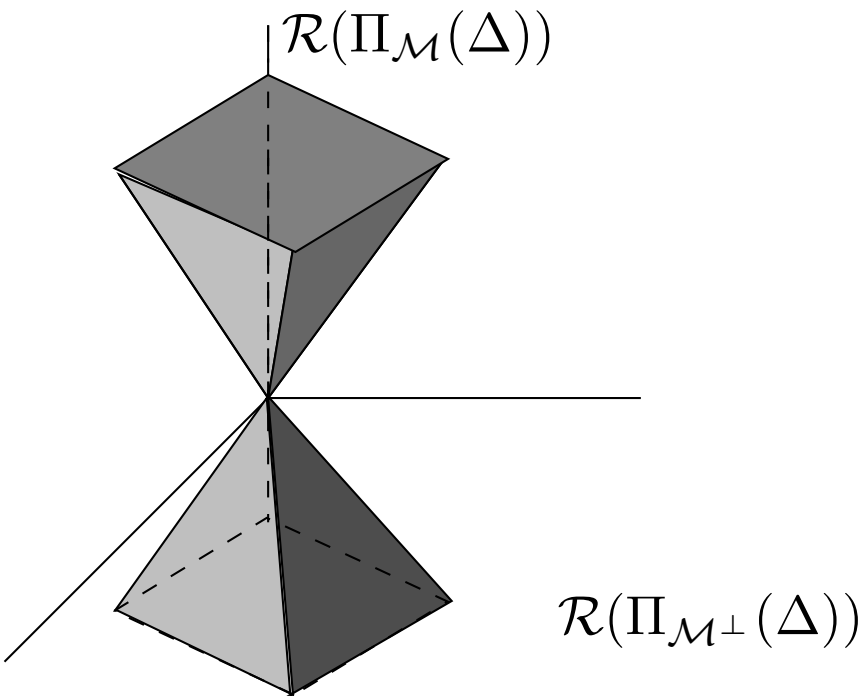
Regularizer $\mathcal{R}$ decomposes across $(\mathcal{M}, \mathcal{M}^{\perp})$ if

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

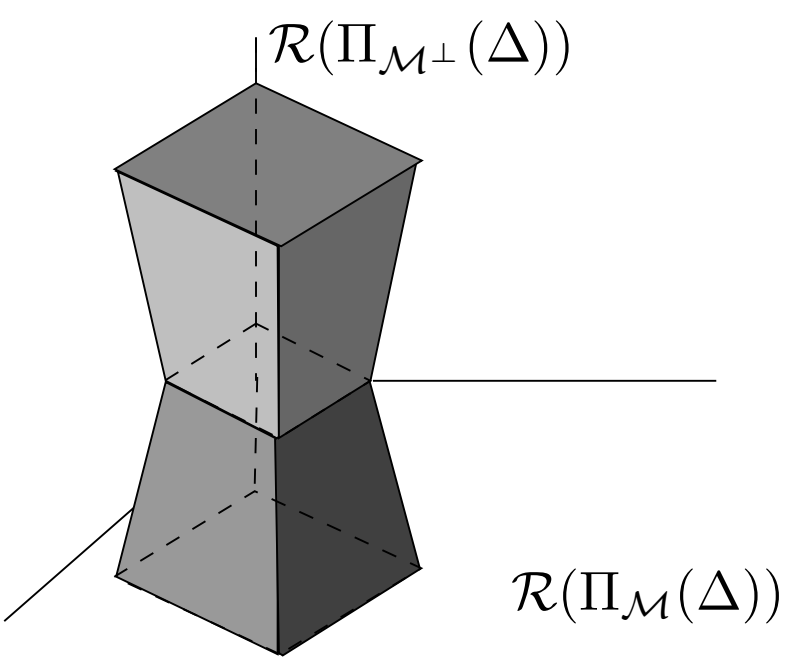
Related definitions:

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. $\mathcal{M}$. 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

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

Example: Sparse vectors and ℓ_1-regularization

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

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

- decomposability of ℓ_1-norm:

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

- 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

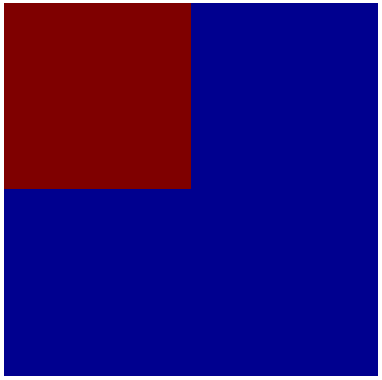
- for each pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{d_1}$ and $V \subseteq \mathbb{R}^{d_2}$:

$$\begin{aligned} \mathcal{M}(U,V) &:= \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{row}(\Theta) \subseteq V, \operatorname{col}(\Theta) \subseteq U\right\} \\ \widetilde{\mathcal{M}}^{\perp}(U,V) &:= \left\{\Gamma \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{row}(\Gamma) \subseteq V^{\perp}, \operatorname{col}(\Gamma) \subseteq U^{\perp}\right\}. \end{aligned}$$

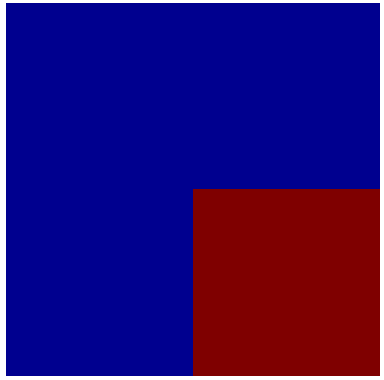
Example: Low-rank matrices and nuclear norm

- for each pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{d_1}$ and $V \subseteq \mathbb{R}^{d_2}$:

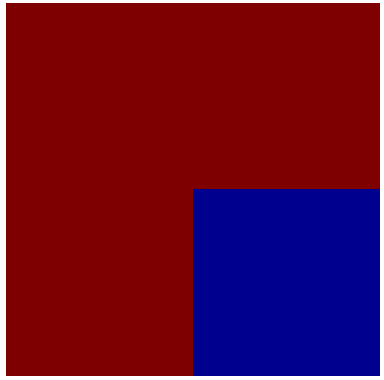
$$\begin{aligned} \mathcal{M}(U,V) &:= \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{row}(\Theta) \subseteq V, \operatorname{col}(\Theta) \subseteq U\right\} \\ \widetilde{\mathcal{M}}^{\perp}(U,V) &:= \left\{\Gamma \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{row}(\Gamma) \subseteq V^{\perp}, \operatorname{col}(\Gamma) \subseteq U^{\perp}\right\}. \end{aligned}$$



(a) $\Theta \in \mathcal{M}$



(b) $\Gamma \in \widetilde{\mathcal{M}}^{\perp}$

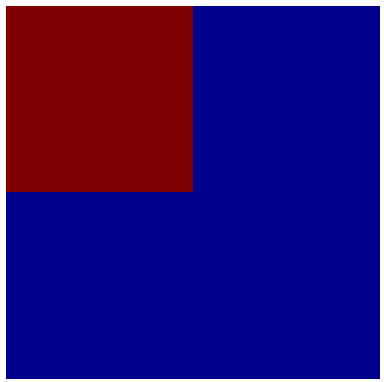


(c) $\Sigma \in \widetilde{\mathcal{M}}$

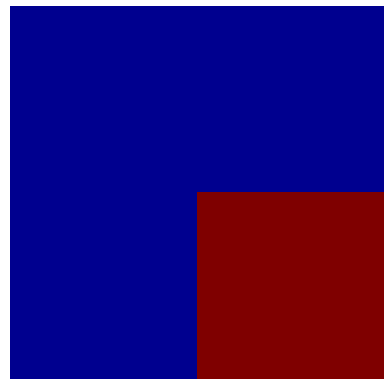
Example: Low-rank matrices and nuclear norm

- for each pair of r-dimensional subspaces $U \subseteq \mathbb{R}^{d_1}$ and $V \subseteq \mathbb{R}^{d_2}$:

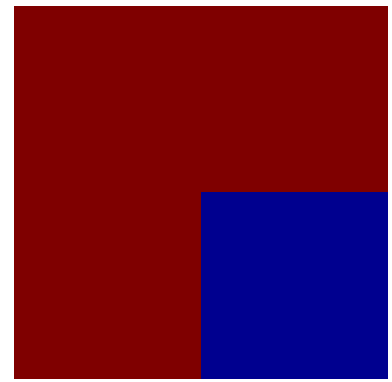
$$\begin{aligned} \mathcal{M}(U,V) &:= \left\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{row}(\Theta) \subseteq V, \operatorname{col}(\Theta) \subseteq U\right\} \\ \widetilde{\mathcal{M}}^{\perp}(U,V) &:= \left\{\Gamma \in \mathbb{R}^{d_1 \times d_2} \mid \operatorname{row}(\Gamma) \subseteq V^{\perp}, \operatorname{col}(\Gamma) \subseteq U^{\perp}\right\}. \end{aligned}$$



(a) $\Theta \in \mathcal{M}$



(b) $\Gamma \in \widetilde{\mathcal{M}}^{\perp}$



(c) $\Sigma \in \widetilde{\mathcal{M}}$

- by construction, $\Theta^T \Gamma = 0$ for all $\Theta \in \mathcal{M}(U,V)$ and $\Gamma \in \widetilde{\mathcal{M}}^{\perp}(U,V)$
- decomposability of nuclear norm $|||\Theta|||_1 = \sum_{j=1}^{\min\{d_1,d_2\}} \sigma_j(\Theta)$:

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

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 $\mathcal{R}$.

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 $\theta^* \in \mathcal{M}$, and $\Psi^2(\mathcal{M})\tau_n^2 < 1$. 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

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

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

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

Theorem (Oracle version)

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

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

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

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: Group-structured regularizers

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

G_1 G_2 G_3

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

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

- 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*

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

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

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

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

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

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_2 regularization: Group norm with $\nu = 2$

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

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*

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

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 $\nu = \infty$

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

Is adaptive estimation possible?

Consider a group-sparse problem with:

- $|\mathcal{G}|$ 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:

- $|\mathcal{G}|$ 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

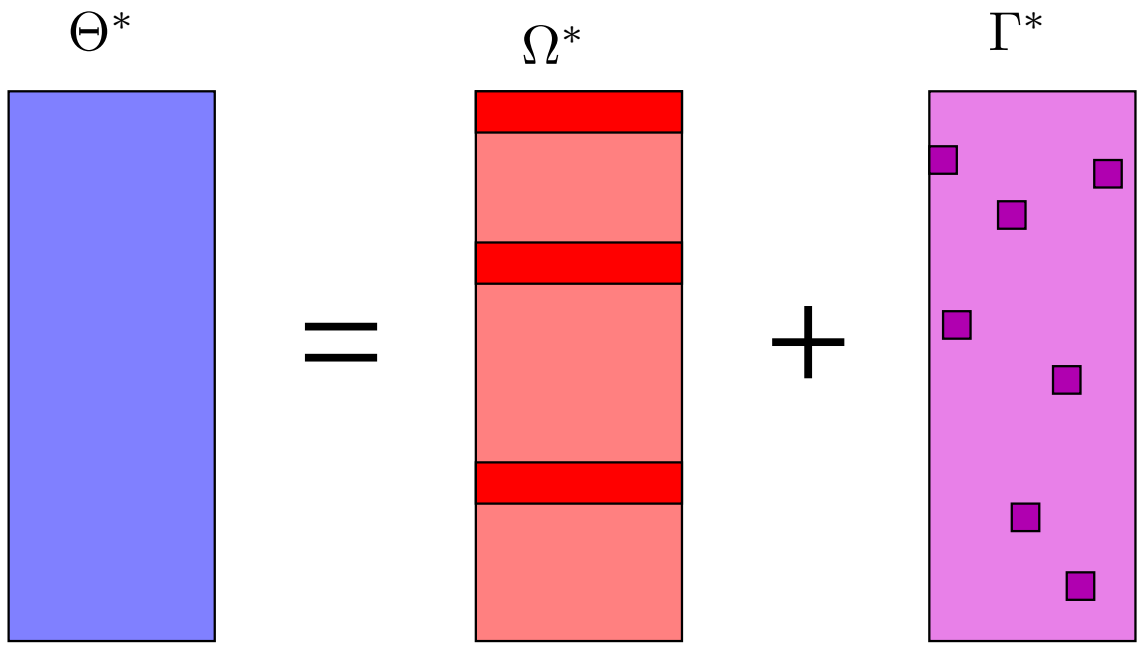
$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{|\mathcal{S}_{\mathcal{G}}|\, T\, \log(|\mathcal{G}|m)}{n}.$$

Natural question:

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

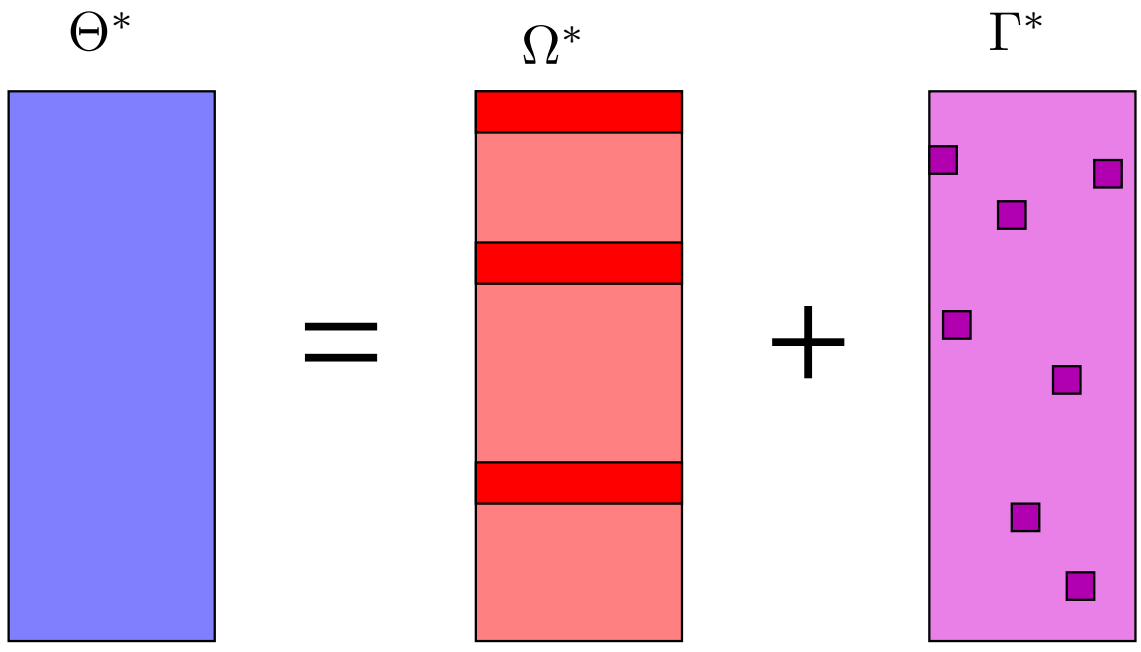
Answer: Overlap group Lasso

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.

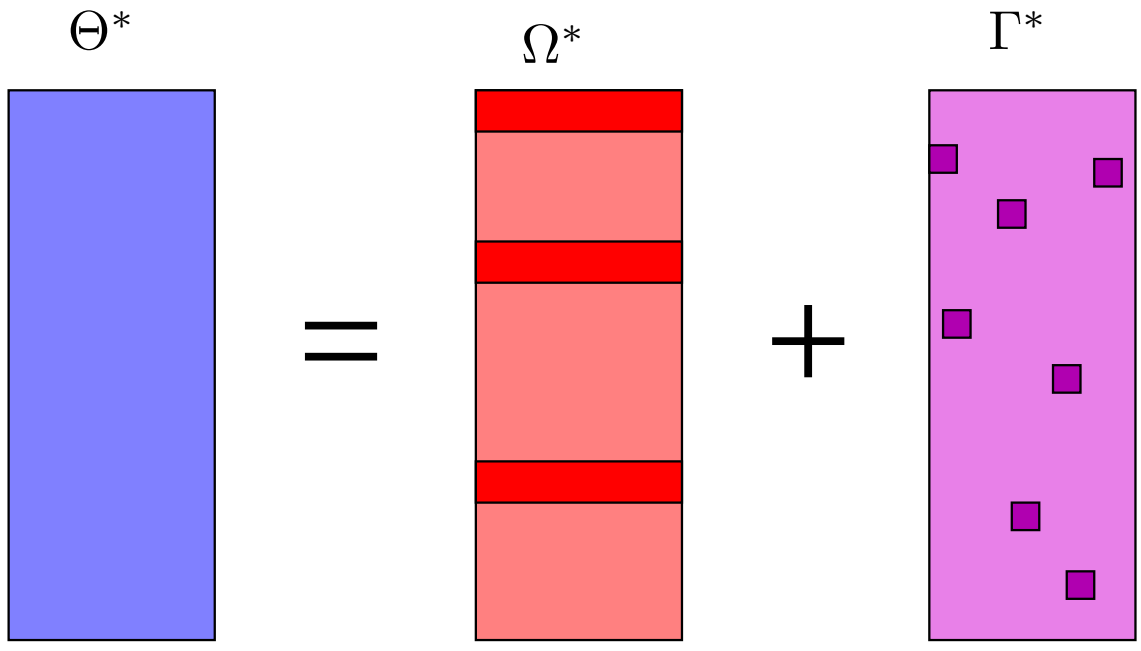


Define new norm on matrix space:

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

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 $\lambda_n = 4\sigma\sqrt{\frac{\log d}{n}}$, any optimal solution satisfies (w.h.p.)

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

Example: Low-rank matrices and nuclear norm

- low-rank matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ 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):

$$\widehat{\Theta} \; \in \; \arg\min_{\Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\!\langle X_i, \, \Theta \rangle\!\rangle)^2 + \lambda_n \underbrace{\sum_{j=1}^{\min\{d_1, d_2\}} \gamma_j(\Theta)}_{|\!|\!|\Theta|\!|\!|_1} \right\}$$

Example: Low-rank matrices and nuclear norm

- low-rank matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ 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):

$$\widehat{\Theta} \; \in \; \arg\min_{\Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\langle X_i, \, \Theta \rangle\rangle)^2 + \lambda_n \underbrace{\sum_{j=1}^{\min\{d_1, d_2\}} \gamma_j(\Theta)}_{|||\Theta|||_1} \right\}$$

- 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:

$$\mathbb{B}_q(R_q) = \big\{\Theta^* \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j=1}^{\min\{d_1, d_2\}} |\sigma_j(\Theta^*)|^q \leq R_q\big\}.$$

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:

$$\mathbb{B}_q(R_q) = \{\Theta^* \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j=1}^{\min\{d_1,d_2\}} |\sigma_j(\Theta^*)|^q \leq R_q\}.$$

Corollary (Negahban & W., 2011)

With regularization parameter $\lambda_n \geq 16\sigma(\sqrt{\frac{d_1}{n}} + \sqrt{\frac{d_2}{n}})$, *we have w.h.p.*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \quad \leq \quad c_0 \frac{R_q}{\gamma_\ell^2} \left(\frac{\sigma^2 (d_1 + d_2)}{n}\right)^{1-\frac{q}{2}}$$

Rates for (near) low-rank estimation

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

$$\mathbb{B}_q(R_q) = \big\{ \Theta^* \in \mathbb{R}^{d_1 \times d_2} \mid \sum_{j=1}^{\min\{d_1,d_2\}} |\sigma_j(\Theta^*)|^q \leq R_q \big\}.$$

Corollary (Negahban & W., 2011)

With regularization parameter $\lambda_n \geq 16\sigma\big(\sqrt{\frac{d_1}{n}} + \sqrt{\frac{d_2}{n}}\big)$, *we have w.h.p.*

$$|||\widehat{\Theta} - \Theta^*|||_F^2 \quad \leq \quad c_0 \, \frac{R_q}{\gamma_\ell^2} \left(\frac{\sigma^2 \, (d_1 + d_2)}{n} \right)^{1 - \frac{q}{2}}$$

- 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

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

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

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

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

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

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

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

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

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

3. 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.

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