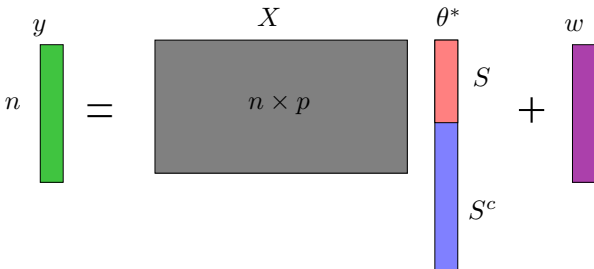


A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers

Sahand Negahban, UC Berkeley
Pradeep Ravikumar, UT Austin
Martin Wainwright, UC Berkeley
Bin Yu, UC Berkeley

NIPS Conference

Example: Sparse regression



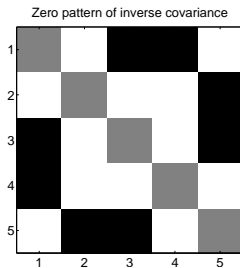
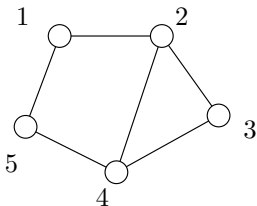
Set-up: noisy observations $y = X\theta^* + w$ with sparse θ^*

Estimator: Lasso program

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

Some past work: Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; Meinshausen/Buhlmann, 2005; Candes/Tao, 2005; Donoho, 2005; Haupt & Nowak, 2006; Zhao/Yu, 2006; Wainwright, 2006; Zou, 2006; Koltchinskii, 2007; Meinshausen/Yu, 2007; Tsybakov et al., 2008

Example: Structured inverse covariance matrices



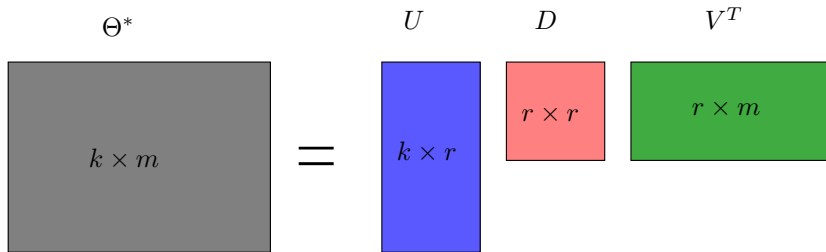
Set-up: Samples from random vector with sparse inverse covariance Θ^* .

Estimator:

$$\hat{\Theta} \in \arg \min_{\Theta} \left\langle \frac{1}{n} \sum_{i=1}^n x_i x_i^T, \Theta \right\rangle - \log \det(\Theta) + \lambda_n \sum_{j=1}^p \|\Theta_j\|_q$$

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

Example: Low-rank matrix approximation



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

Estimator:

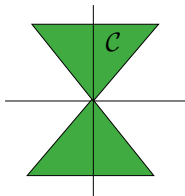
$$\hat{\Theta} \in \arg \min_{\Theta} \frac{1}{n} \sum_{i=1}^n (y_i - \langle X_i, \Theta \rangle)^2 + \lambda_n \sum_{j=1}^{\min\{k, m\}} \sigma_j(\Theta)$$

Some past work: Frieze et al., 1998; Achilioptas & McSherry, 2001; Srebro et al., 2004; Drineas et al., 2005; Rudelson & Vershynin, 2006; Recht et al., 2007; Bach, 2008; Meka et al., 2008; Candes & Tao, 2009; Keshavan et al., 2009

Important properties of regularizer/loss

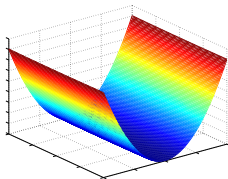
1 Decomposability of regularizer

- ▶ vectors $u \in A$ and $v \in B \Rightarrow r(u + v) = r(u) + r(v)$
- ▶ constrains error $\Delta = \hat{\theta} - \theta^*$ to smaller set \mathcal{C}



2 Restricted strong convexity:

- ▶ loss functions not strictly convex in high-dimensions
- ▶ require “curvature” only for directions $\Delta \in \mathcal{C}$
- ▶ loss function $\mathcal{L}_n(\theta) := \mathcal{L}_n(\theta; \mathcal{X}_1^n)$ satisfies



$$\underbrace{\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*)}_{\text{Excess loss}} - \underbrace{\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle}_{\text{score function}} \geq \gamma(\mathcal{L}) \underbrace{d^2(\Delta)}_{\text{squared error}} \quad \text{for all } \Delta \in \mathcal{C}.$$

Main theorem

Quantities that control rates:

- restricted strong convexity parameter: $\gamma(\mathcal{L})$
- dual norm of regularizer: $r^*(v) := \sup_{r(u)=1} \langle v, u \rangle$.
- optimal subspace const.: $\Psi(A) = \min \{c \in \mathbb{R} \mid r(\theta) \leq c d(\theta) \text{ for all } \theta \in A\}$.

Theorem

With regularization constant $\lambda_n \geq 2r^*(\nabla \mathcal{L}(\theta^*; \mathcal{X}_1^n))$, then any solution $\hat{\theta}$ satisfies

$$d(\hat{\theta} - \theta^*) \leq \frac{1}{\gamma(\mathcal{L})} [\Psi(B^\perp) \lambda_n].$$

Assumptions:

- θ^* belongs to a subspace A
- regularizer r decomposable over subspace pair (A, B)
- loss obeys restricted strong convexity with parameter $\gamma(\mathcal{L}) > 0$

Application: Linear regression (hard sparsity)

- RSC reduces to lower bound on restricted eigenvalues of $X^T X$
- 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 $\lambda_n \geq 2 \left\| \frac{X^T \varepsilon}{n} \right\|_\infty$, then any Lasso solution satisfies $\|\hat{\theta} - \theta^*\|_2 \leq \frac{1}{\gamma(\mathcal{L})} \sqrt{k} \lambda_n$.

Some stochastic instances: recover known results

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

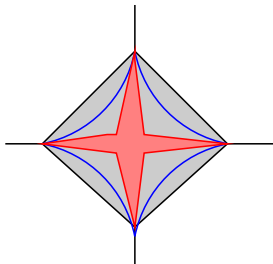
$$\left\| \frac{X^T \varepsilon}{n} \right\|_\infty \leq \sqrt{\frac{2\sigma^2 \log p}{n}} \quad \text{w.h.p.} \implies \|\hat{\theta} - \theta^*\|_2 \leq \frac{8\sigma}{\gamma(\mathcal{L})} \sqrt{\frac{k \log p}{n}}.$$

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

Application: Linear regression (weak sparsity)

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

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



Corollary

Under RSC, then any Lasso solution satisfies (w.h.p.)

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \mathcal{O}\left[\sigma^2 R_q \left(\frac{\log p}{n}\right)^{1-q/2}\right].$$

- new result; rate known to be minimax optimal (Raskutti et al., 2009)

Multivariate regression with block regularizers

$$\begin{array}{c} Y \\ n \\ \color{green}{\boxed{}} \\ m \end{array} = \begin{array}{c} X \\ n \times p \\ \color{gray}{\boxed{}} \end{array} = \begin{array}{c} \Theta^* \\ \color{red}{S} \\ \color{blue}{S^c} \\ p \\ m \end{array} + \begin{array}{c} W \\ m \\ \color{purple}{\boxed{}} \\ m \end{array}$$

- ℓ_1/ℓ_q -regularized group Lasso: with $\lambda_n \geq 2 \left\| \frac{X^T W}{n} \right\|_{\infty, \tilde{q}}$ where $1/q + 1/\tilde{q} = 1$

$$\hat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{p \times p}} \left\{ \frac{1}{2n} \|Y - X\Theta\|_F^2 + \lambda_n \|\Theta\|_{1,q} \right\}.$$

Corollary

Say Θ^* is supported on $|S| = s$ rows, X satisfies *RSC* and $W_{ij} \sim N(0, \sigma^2)$. Then we have $\|\hat{\Theta} - \Theta^*\|_F \leq \frac{2}{\gamma(\mathcal{L})} \Psi_q(S) \lambda_n$ where

$$\Psi_q(S) = \begin{cases} m^{1/q-1/2} \sqrt{s} & \text{if } q \in [1, 2). \\ \sqrt{s} & \text{if } q \geq 2. \end{cases}$$

Multivariate regression with block regularizers

$$\begin{matrix} Y \\ n \\ m \end{matrix} = \begin{matrix} X \\ n \times p \end{matrix} + \begin{matrix} \Theta^* \\ S \\ S^c \\ p \\ m \end{matrix} + \begin{matrix} W \\ n \\ m \end{matrix}$$

Effect of varying $q \in [1, \infty]$:

- for $q = 1$, problem reduces ordinary Lasso with pm parameters and sparsity sm :

$$\|\hat{\Theta} - \Theta^*\|_F \leq \mathcal{O}\left(\sqrt{\frac{sm \log(pm)}{n}}\right)$$

- for $q = 2$, rate decouples into term terms:

$$\|\hat{\Theta} - \Theta^*\|_F \leq \mathcal{O}\left(\underbrace{\sqrt{\frac{s \log p}{n}}}_{\text{Search term (find } s \text{ rows)}} + \underbrace{\sqrt{\frac{sm}{n}}}_{\text{Estimate } sm \text{ parameters}}\right)$$

Search term (find s rows)

Estimate sm parameters

- similar rates for $q = 2$: Lounici et al. (2009) and Huang and Zhang (2009)

Application: Low-rank matrices and nuclear norm

- low-rank matrix $\Theta^* \in \mathbb{R}^{k \times m}$ with rank $r \leq \min\{k, m\}$
- noisy/partial observations of the form

$$y_i = \langle X_i, \Theta^* \rangle + \varepsilon_i, \quad i = 1, \dots, n, \quad \varepsilon_i \sim N(0, \sigma^2).$$

Corollary

With regularization parameter $\lambda_n \geq 16\sigma \left(\sqrt{\frac{k}{n}} + \sqrt{\frac{m}{n}} \right)$, we have w.h.p.

$$\|\hat{\Theta} - \Theta^*\|_F \leq \frac{32\sigma}{\gamma(\mathcal{L})} \left[\sqrt{\frac{r k}{n}} + \sqrt{\frac{r m}{n}} \right].$$

- for a rank r matrix M , we have $\|M\|_1 \leq \sqrt{r} \|M\|_F$
- solve nuclear norm regularized program with $\lambda_n \geq \frac{2}{n} \left\| \sum_{i=1}^n X_i \varepsilon_i \right\|_2$

Summary

- unified approach to 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 d
 - ▶ restricted strong convexity constant
- recovered some known results as corollaries:
 - ▶ Lasso with exact sparsity
 - ▶ multivariate group Lasso
 - ▶ inverse covariance matrix estimation
- derived new results on:
 - ▶ low-rank matrix estimation
 - ▶ “approximately” sparse models
 - ▶ other models?