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

Loss functions and regularization

- **Model class:** parameter space $\Omega \subset \mathbb{R}^p$, and set of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$

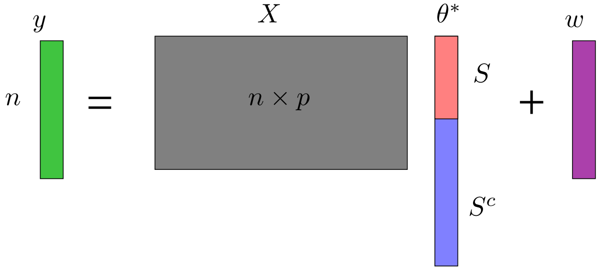
- **Data:** samples $\mathcal{X}_1^n = (x_i, y_i)$, $i = 1, \ldots, n$ are drawn from unknown $\mathbb{P}_{\theta^*}$

- **Estimation:** Minimize loss function plus regularization term:

$$\underbrace{\widehat{\theta}}_{\text{Estimate}} \in \arg\min_{\theta \in \mathbb{R}^p} \Big\{ \underbrace{\mathcal{L}_n(\theta; \mathcal{X}_1^n)}_{\text{Loss function}} + \lambda_n \underbrace{r(\theta)}_{\text{Regularizer}} \Big\}.$$

- **Analysis:** Bound error $d(\widehat{\theta} - \theta^*)$ under high-dimensional scaling $(n, p) \to +\infty$.

Example: Sparse regression



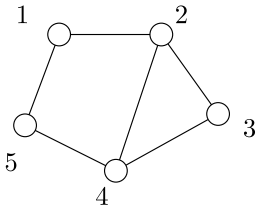
Set-up: noisy observations $y = X\theta^* + w$ with sparse θ^*

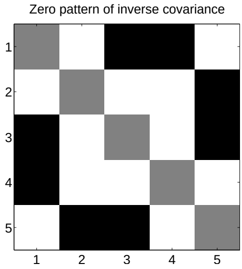
Estimator: Lasso program

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

<u>Some past work:</u> Tibshirani, 1996; Chen et al., 1998; Donoho/Xuo, 2001; Tropp, 2004; Fuchs, 2004; 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





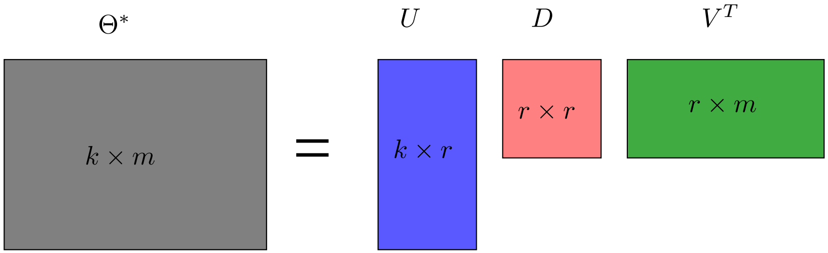
Set-up: Samples from random vector with sparse inverse covariance Θ^*.

Estimator:

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

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 rank $r \ll \min\{k, m\}$.

Estimator:

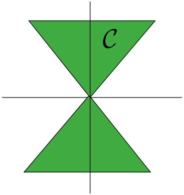
$$\widehat{\Theta} \in \arg\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle\langle X_i,\ \Theta \rangle\rangle)^2 + \lambda_n \sum_{j=1}^{\min\{k,m\}} \sigma_j(\Theta)$$

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

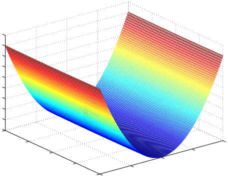
1 Decomposability of regularizer

- vectors $u \in A$ and $v \in B \Rightarrow$ $r(u+v) = r(u) + r(v)$
- constrains error $\Delta = \widehat{\theta} - \theta^*$ 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

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

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

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

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 $\lambda_n \geq 2\|\frac{X^T\varepsilon}{n}\|_\infty$, then any Lasso solution satisfies $\|\widehat{\theta} - \theta^*\|_2 \leq \frac{1}{\gamma(\mathcal{L})}\sqrt{k}\,\lambda_n$.

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

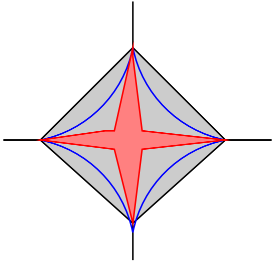
$$\|\frac{X^T\varepsilon}{n}\|_\infty \leq \sqrt{\frac{2\sigma^2 \log p}{n}} \quad \text{w.h.p.} \implies \|\widehat{\theta} - \theta^*\|_2 \leq \frac{8\sigma}{\gamma(\mathcal{L})}\sqrt{\frac{k\,\log p}{n}}.$$

(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"

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



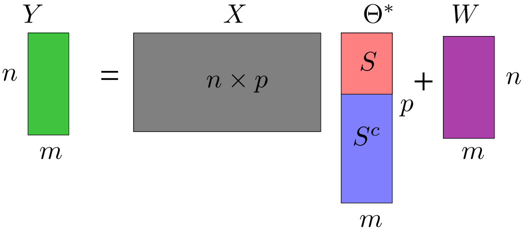
Corollary

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

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

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

Multivariate regression with block regularizers



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

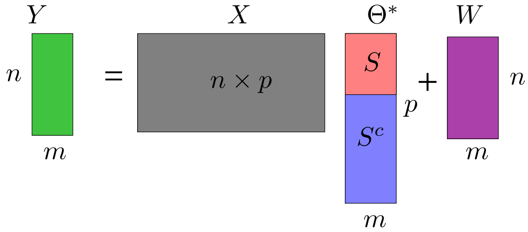
$$\widehat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{p \times p}} \left\{ \frac{1}{2n} \|Y - X\Theta\|_F^2 + \lambda_n \|\Theta\|_{1,q} \right\}.$$

Corollary

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



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

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

$$\|\widehat{\Theta} - \Theta^*\|_F \quad \leq \quad \mathcal{O}\Big(\sqrt{\frac{sm \log(pm)}{n}}\Big)$$

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

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

- 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

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

Corollary

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

$$|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F \quad \leq \quad \frac{32\sigma}{\gamma(\mathcal{L})} \Big[\sqrt{\frac{r \; k}{n}} + \sqrt{\frac{r \; m}{n}} \Big].$$

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

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?