Regression Performance of Group Lasso for Arbitrary Design Matrices

Marco F. Duarte

Joint work with Waheed U. Bajwa and Robert Calderbank
Linear Regression and the Lasso

- One regression variable per unit-norm column of $X$
- Modeling error assumed to be i.i.d. $z \sim \mathcal{N}(0, \sigma^2 I)$
- Lasso: obtain sparse regression coefficient vector as

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2 + 2\lambda\sigma \|\beta\|_1$$

[Tibshirani, 1996]
Group Linear Regression

- Correlated variables grouped in $m$ column submatrices
- Group Lasso: use mixed norm on coefficient vector

\[
\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^{pm}} \| y - X\beta \|_2 + 2\lambda \sigma \sqrt{m} \| \beta \|_{2,1}
\]

where $\| \beta \|_{2,1} = \sum_{i=1}^{p} \| \beta_i \|_2$

[Yuan and Lin, 2006]
Existing Performance Guarantees for Group Lasso

• **Asymptotic** convergence for linear regression, coefficient estimation over *random design matrix*
  [Bach 2008][Meier, van de Geer, Bühlmann 2008]

• **Asymptotic** convergence for linear regression, coefficient estimation, model selection over random noise
  [Liu and Zhang 2009][Nardi and Rinaldo 2010]

• Non-asymptotic convergence of linear regression, coefficient estimation over random noise via
  *combinatorially complex matrix conditions*
  [Chesneau and Hebiri][Huang and Zhang 2010]

• **Today**: non-asymptotic convergence for linear regression over random noise via *simple matrix conditions*
Recent Analytical Tools for Lasso

- **Probabilistic model** on regression coefficient vector:
  - support $I \subseteq \{1, \ldots, p\}$ of $\beta_I$ selected uniformly at random
  - signs of $k$ nonzero entries $\beta$ are i.i.d. and equally likely $\pm 1$

- **Simple metrics** on design matrix
  - spectral norm $\|X\|_2$
  - worst-case coherence $\mu(X) = \max_{1 \leq i \neq j \leq p} |\langle X_i, X_j \rangle|$
Theorem: Assume that
• \( k \leq \frac{C_0 p}{\|X\|_2^2 \log p} \) and
• \( \mu(X) \leq C_1 / \log p \).

If \( \lambda = \sqrt{2 \log p} \) and \( z \sim \mathcal{N}(0, \sigma^2 I) \), then the output of lasso obeys we can guarantee that

\[
\|X\beta - X\hat{\beta}\|_2^2 \leq Ck\sigma^2 \log p
\]

with probability at least \( 1 - \mathcal{O}(p^{-1}) \).

[Candès, Plan 2010]
New Analytical Tools for Group Lasso

- **New definition** of sign patterns for grouped regression coefficient vector \( \beta = [\beta_1 \beta_2 \ldots \beta_p] \)

  \[
  \overline{\text{sign}}(\beta_i) = \beta_i / \|\beta_i\|_2
  \]

- **Probabilistic model** on regression coefficient vector that accounts for correlations within groups:
  - active groups \( I \subseteq \{1, \ldots, p\} \) of \( \beta \) selected uniformly at random
  - group signs of nonzero groups of \( \beta \) are statistically independent:

  \[
  \mathbb{P}(\bigcup_{i \in I} \overline{\text{sign}}(\beta_i) \in A_i) = \prod_{i \in I} \mathbb{P}(\overline{\text{sign}}(\beta_i) \in A_i)
  \]

  - nonzero regression coefficients have zero median:

  \[
  \mathbb{E}(\overline{\text{sign}}(\beta)) = 0
  \]
Simple Metrics on Design Matrix

• Spectral norm $\|X\|_2$

• Worst-case coherence:

$$\mu(X) = \max_{1 \leq i \neq i', 1 \leq j \neq j' \leq m} |\langle X_{i,j}, X_{i',j'} \rangle|$$

• Worst-case *block coherence:*

$$\mu_B(X) = \max \left\{ \max_{1 \leq i \neq i' \leq p} \|X_i^T X_{i'}\|_2, \max_{1 \leq i \leq p} \|X_i^T X_i - I\|_2 \right\}$$

See also [Eldar, Rauhut 2010]
Theorem: Assume that

- \( k \leq \frac{C_0 p}{\|X\|_2^2 \log(pm)} \),
- \( \mu(X) \leq 1/m \), and
- \( \mu_B(X) \leq C_1 / \log(pm) \).

If \( \lambda = \sqrt{2 \log(pm)} \) and \( z \sim \mathcal{N}(0, \sigma^2 I) \), then the output of group lasso obeys

\[
\|X\beta - X\hat{\beta}\|_2^2 \leq Cm k \sigma^2 \log(pm)
\]

with probability at least \( 1 - O((pm)^{-1}) \).

Note that for \( m=1 \), group lasso is same as lasso, and we obtain the result of [Candès, Plan 2010].
**Lemma** [Duarte, Bajwa, Calderbank 2010]

Define i.i.d. Bernoulli random variables $\delta_1, \ldots, \delta_p$ with parameter $\delta = k/p$ and form a block submatrix $X_{I'} = [X_i : \delta_i = 1]$. Then for $q = 2\log(pm)$, we have the bound

$$[\mathbb{E}\|X_{I'}^*X_{I'} - I\|_2^q]^{1/q} \leq 20\mu_B(X)\log(pm) + \delta\|X\|_2^2$$

$$+ 9\sqrt{\delta\log(pm)(1 + (m - 1)\mu(X))}\|X\|_2^2$$

- Generalizes results on conditioning of random subdictionaries [Tropp 2008] to grouped submatrices
- Theorem’s conditions on $\mu_B(X), \mu(X), \|X\|_2$ provide $O(1)$ bounds on $[\mathbb{E}\|X_{I'}^*X_{I'} - I\|_2^q]^{1/q}$
Proof Sketch

(Well) conditioning in expectation, combined with distribution on $z$, imply that with the given probability these three properties hold simultaneously:

- **invertibility**: $X_I^* X_I$ is invertible and $\| (X_I^* X_I)^{-1} \|_2 \leq 2$
- **orthogonality**: $\| X^* z \|_{2,\infty} \leq \sqrt{2m}\lambda$
- **complementary size**:

$$2\lambda \sqrt{m} \| X_{I^c}^* X_I (X_I^* X_I)^{-1} \text{sign}(\beta_I) \|_{2,\infty}$$

$$+ \| X_{I^c}^* X_I (X_I^* X_I)^{-1} X_I^* z \|_{2,\infty} \leq (2 - \sqrt{2})\lambda \sqrt{m}$$

where $\| \beta \|_{2,\infty} = \max_{1 \leq i \leq p} \| \beta_i \|_2$

(similar to [Candès, Plan 2010])
• \( m \) sparse **correlated** vectors \( B = [\beta_1 \beta_2 \ldots \beta_m] \in \mathbb{R}^{p \times m} \)

• Vectors \( \beta_i \) share **common support**

• **Single** design matrix \( X \)

• Observation model \( Y = XB + Z \), error \( Z \in \mathbb{R}^{n \times m} \)

• MMV problem can alternatively be expressed as group sparse linear regression:

\[
y' = \text{vect} \left( Y^T \right) \quad \beta' = \text{vect} \left( B^T \right) \quad z' = \text{vect} \left( Z^T \right) \\
X' = X^T \otimes I \\
y' = X' \beta' + z'
\]
Group Lasso and Multiple Measurement Vectors

\[ y' = \text{vect} \left( Y^T \right) \quad \beta' = \text{vect} \left( B^T \right) \quad z' = \text{vect} \left( Z^T \right) \]

\[ X' = X^T \otimes I \]

\[ y' = X' \beta' + z' \]

Guarantees available for many MMV recovery algorithms:

- rely on random design matrix,
- focus on asymptotic behavior,
- focus on coefficient estimation or model selection,
- rely on combinatorially complex matrix metrics, or
- apply only for noiseless model

Recent tools for average case analysis of linear regression via lasso can be extended to group lasso.

New probability model that captures correlations present within each group of predictor variables.

Can apply this probability model to standard lasso: group lasso relaxes lasso’s requirement from \( \mu_B(X) \leq C_1/m \log(pm) \) to \( \mu_B(X) \leq C_1/\log(pm) \).

Extending guarantees to model selection, coefficient estimation for grouped variables.

http://www.cs.duke.edu/~mduarte