Support Stability of Group Lasso

Mohammad Golbabaee and Gabriel Peyré
Centre de Recherche en Mathmatiques de la Décision (CEREMADE), CNRS and Université Paris-Dauphine
E-mail:{golbabaee,peyre}@ceremade.dauphine.fr

***Abstract*—This paper proposes a novel signal-dependent criterion to guarantee the stability of support recovery using group-Lasso regularization. This criterion ensures that, when the signal-to-noise ratio is large enough, $\ell_1 - \ell_2$ block sparsity regularization recovers a signal with the same block support as the original signal. Consequently, this implies a linear convergence of the ℓ_2 recovery error when the noise tends to zero. In the noiseless case, this criterion guarantees that one recovers exactly the original signal.**

I. Group Lasso Regularization

Given a set of non-overlapping blocks $\mathcal{B}$ describing a partition $\{1,\ldots,N\} = \bigcup_{b\in\mathcal{B}} b$, the block norms $\ell_1 - \ell_2$ and $\ell_\infty - \ell_2$ of $x \in \mathbb{R}^N$ are defined as

$$\|x\|_{1,2} \triangleq \sum_{b\in\mathcal{B}} \|x_b\|_2 \quad \text{and} \quad \|x\|_{\infty,2} \triangleq \max_{b\in\mathcal{B}} \|x_b\|_2. \tag{1}$$

We consider the problem of recovering $x^0 \in \mathbb{R}^N$ from a set of linear and noisy observations $y = \Phi x^0 + w$, where $\Phi \in \mathbb{R}^{d\times N}$ is the design operator, and $w \in \mathbb{R}^d$ is the noise vector. The design operator Φ is typically ill-posed and thus, recovering a precise approximation of x_0 from y is a challenging inverse problem. The block sparsity structure of the original vector appears to be beneficial for obtaining a robust estimation. Following for instance [1], a group Lasso estimator is proposed by solving the following convex problem:

$$(\mathcal{P}_\lambda(y)) \quad x^\star \in \operatorname*{argmin}_x \frac{1}{2}\|y - \Phi x\|_2^2 + \lambda\|x\|_{1,2}.$$

The main objective of this article is to analyze the performance of $(\mathcal{P}_\lambda(y))$ for a deterministic (non-probabilistic) setup of x^0, Φ and w.

II. Support Stability

a) Main Contribution: Let us first fix some notations: we denote by $I \triangleq \{b \in \mathcal{B} \setminus x_b^0 \neq 0\}$ the support set of x^0, where $x_b^0 = (x_i^0)_{i\in b}$. We denote by $T = \min_{b\in I} \|x_b^0\|_2 > 0$ the signal level. Assuming that Φ_I is injective, the *identifiably criterion* associated to x^0 reads

$$\mathbf{IC}(x_0) = \|\Phi_{I^c}^* \Phi_I^{+,*} \mathcal{N}(x_{0,I})\|_{\infty,2} \text{ where } \mathcal{N}(x_I) = \left(\frac{x_b}{\|x_b\|}\right)_{b\in I},$$

$\Phi_I^+ = (\Phi_I^*\Phi_I)^{-1}\Phi_I^*$, and I^c denotes the co-support. The following theorem presents the main contribution of this article:

Theorem 1: Suppose Φ_I is an injective operator and $\mathbf{IC}(x^0) < 1$. $\exists c_I, \tilde{c}_I > 0$ *such that, if $\|w\|_2/T < \tilde{c}_I/c_I$ and $c_I\|w\|_2 < \lambda < \tilde{c}_I T$, the solution $x^\star$ of $\mathcal{P}_\lambda(y)$ is unique and it is exactly supported on I. In addition, $\|x^\star - x^0\|_2 \leqslant C_I\|w\|_2 + \tilde{C}_I\lambda$ for some constants $C_I, \tilde{C}_I > 0$.*

In plain words, this theorem asserts that for a properly chosen λ proportional to the noise level, the condition $\mathbf{IC}(x^0) < 1$ ensures both exact support recovery (i.e. $x^\star$ shares the same support as x^0), and a linear convergence of the ℓ_2 recovery error i.e. $\|x^0 - x^\star\|_2 = O(\|w\|_2)$. This implies exact signal recovery $x^\star = x^0$ when $w = 0$ and $\lambda = 0^+$, as previously shown in [2].

b) Previous and Related Works: Theorem 1 is proved by Fuchs [3] in the special case of the Lasso i.e. the ℓ_1 regularization (having blocks of size 1). Eldar and Rauhut [2] derived the same identifiably criterion, however, ensuring exact block sparse recovery only in the noiseless case i.e. $w = 0$ and $\lambda = 0^+$. Bach [4] showed that $\mathbf{IC}(x^0) < 1$ ensures consistency of the group lasso when Φ is injective, which corresponds to the convergence of $x^\star$ to x_0 when $\lambda \sim \|w\| \to 0^+$. We extend these results to the inverse problem setting where Φ might not be injective and we provide non-asymptotic bounds on the signal-to-noise ratio to ensure support identifiability.

III. Sketch of the Proof

We construct $\hat{x}$ with $\hat{x}_{I^c} = 0$ and

$$\hat{x}_I = \operatorname*{argmin}_{x_I} \mathcal{F}(x_I) \text{ where } \mathcal{F}(x_I) = \frac{1}{2}\|y - \Phi_I x_I\|^2 + \lambda\|x_I\|_{1,2} \tag{2}$$

The proof proceeds by showing that $\hat{x} = x^\star$ is the solution to this problem when λ is well chosen. This requires to show that

$$(C_1) \quad \|\Phi_{I^c}^*(y - \Phi_I\hat{x}_I)\|_{\infty,2} < \lambda,$$

and $(\hat{x})_b \neq 0$ for all $b \in I$. The later is implied by the condition

$$(C_2) \quad \|\hat{x}_I - x_I^0\|_{\infty,2} < T.$$

For an injective Φ_I, the optimality condition i.e., $0 \in \partial\mathcal{F}(\hat{x}_I)$, gives

$$\|\hat{x}_I - x_I^0\|_{\infty,2} \leqslant \|\Phi_I^+ w\|_{\infty,2} + \lambda\|(\Phi_I^*\Phi_I)^{-1}u\|_{\infty,2}, \tag{3}$$

where $u \in \partial(\|\hat{x}_I\|_{1,2})$, implying $\|u\|_{\infty,2} \leqslant 1$. Using (3) we can show that the condition (C_2) is implied by a stronger condition of the form:

$$(C_2') \quad D\varepsilon + E\lambda < T,$$

where $\varepsilon = \|w\|_2$ and $E, D > 0$ are constants. Once (C_2') holds, the optimality condition of (2) implies the following implicit equation:

$$\hat{x}_I = x_I^0 + \Phi_I^+ w - \lambda(\Phi_I^*\Phi_I)^{-1}\mathcal{N}(\hat{x}_I). \tag{4}$$

On the other hand, by inserting the identity (4) in (C_1) and using the bound $\|\mathcal{N}(\hat{x}_I) - \mathcal{N}(x_I^0)\|_{\infty,2} \leqslant \frac{2}{T}\|\hat{x}_I - x_I^0\|_{\infty,2}$, we can show that the condition (C_1) is implied by a stronger condition of the form

$$(C_1') \quad A\varepsilon - (1 - \mathbf{IC}(x^0))\lambda + B\lambda\varepsilon + C\lambda^2 < 0,$$

for some constants $A, B, C, D > 0$. Conditions (C_1') and (C_2') are polynomial constraints that define an admissible set for λ given ε and T . Standard algebraic manipulations show that this admissible set can be bounded in the form announced in Theorem 1.

References

[1] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. of The Roy. Stat. Soc. B*, vol. 68, no. 1, pp. 49–67, 2006.

[2] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[3] J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Transactions on Information Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[4] F. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

Proof of Theorem 1

Mohammad Golbabaee and Gabriel Peyré
Centre de Recherche en Mathmatiques de la Décision (CEREMADE), CNRS and Université Paris-Dauphine
E-mail:{golbabaee,peyre}@ceremade.dauphine.fr

I. Notations

Given a set of non-overlapping blocks $\mathcal{B}$ describing a partition $\{1,\ldots,N\} = \bigcup_{b\in\mathcal{B}} b$, the block norms $\ell_1 - \ell_2$ and $\ell_\infty - \ell_2$ of $x \in \mathbb{R}^N$ are defined as

$$\|x\|_{1,2} \triangleq \sum_{b\in\mathcal{B}} \|x_b\|_2 \quad \text{and} \quad \|x\|_{\infty,2} \triangleq \max_{b\in\mathcal{B}} \|x_b\|_2. \tag{1}$$

Additionally, we define black induced norms for a matrix M as

$$\|M\|_{\infty,2\to\infty,2} \triangleq \max_{\|x\|_{\infty,2}=1} \|Mx\|_{\infty,2} \tag{2}$$

$$\|M\|_{\infty,2\to 2} \triangleq \max_{\|x\|_{2}=1} \|Mx\|_{\infty,2} \tag{3}$$

We denote by $I \triangleq \left\{b \in \mathcal{B} \setminus x_b^0 \neq 0\right\}$ the support set of x^0, where $x_b^0 = (x_i^0)_{i\in b}$. We denote by $T = \min_{b\in I} \|x_b^0\|_2 > 0$ the signal level. Assuming that Φ_I is injective, the *identifiably criterion* associated to x^0 reads

$$\mathbf{IC}(x_0) = \|\Phi_{I^c}^* \Phi_I^{+,*} \mathcal{N}(x_{0,I})\|_{\infty,2} \text{ where } \mathcal{N}(x_I) = \left(\frac{x_b}{\|x_b\|}\right)_{b\in I},$$

$\Phi_I^+ = (\Phi_I^*\Phi_I)^{-1}\Phi_I^*$, and I^c denotes the co-support.

Theorem 1: Suppose Φ_I is an injective operator and $\mathbf{IC}(x^0) < 1$. $\exists c_I, \tilde{c}_I > 0$ such that, if $\|w\|_2/T < \tilde{c}_I/c_I$ and $c_I\|w\|_2 < \lambda < \tilde{c}_I T$, the solution $x^\star$ of $\mathcal{P}_\lambda(y)$ is unique and it is exactly supported on I. In addition, $\|x^\star - x^0\|_2 \leqslant C_I\|w\|_2 + \tilde{C}_I\lambda$ for some constants $C_I, \tilde{C}_I > 0$.

II. Proof

We construct $\hat{x}$ with $\hat{x}_{I^c} = 0$ and

$$\hat{x}_I = \underset{x_I}{\operatorname{argmin}}\ \mathcal{F}(x_I) \quad \text{where} \quad \mathcal{F}(x_I) = \frac{1}{2}\|y - \Phi_I x_I\|^2 + \lambda\|x_I\|_{1,2} \tag{4}$$

The proof proceeds by showing that $\hat{x} = x^\star$ is the solution to this problem when λ is well chosen. This requires to show that

- The elements on the support do not vanish i.e. $(\hat{x})_b \neq 0$ for all $b \in I$. This is implied by the condition
$$(C_2) \quad \|\hat{x}_I - x_I^0\|_{\infty,2} < T.$$
- Uniqueness is implied by the following condition (**need reference**)
$$(C_1) \quad \|\Phi_{I^c}^*(y - \Phi_I\hat{x}_I)\|_{\infty,2} < \lambda,$$

For an injective Φ_I, the optimality condition i.e., $0 \in \partial\mathcal{F}(\hat{x}_I)$, gives

$$\|\hat{x}_I - x_I^0\|_{\infty,2} \leqslant \|\Phi_I^+ w\|_{\infty,2} + \lambda\|(\Phi_I^*\Phi_I)^{-1}u\|_{\infty,2}, \tag{5}$$

for all $u \in \mathbb{R}^{|I|}$ and $\|u\|_{\infty,2} \leqslant 1$. Using (5) we can show that the condition (C_2) is implied by a stronger condition of the form:

$$(C_2') \quad D\varepsilon + E\lambda < T,$$

where $\varepsilon = \|w\|_2$ and $E, D > 0$ are constants. Once (C_2') holds, the optimality condition of (4) implies the following implicit equation:

$$\hat{x}_I = x_I^0 + \Phi_I^+ w - \lambda(\Phi_I^*\Phi_I)^{-1}\mathcal{N}(\hat{x}_I). \tag{6}$$

On the other hand, by inserting the identity (6) in (C_1) and using the bound $\|\mathcal{N}(\hat{x}_I) - \mathcal{N}(x_I^0)\|_{\infty,2} \leqslant \frac{2}{T}\|\hat{x}_I - x_I^0\|_{\infty,2}$, we can show that the condition (C_1) is implied by a stronger condition of the form

$$(C_1') \quad A\varepsilon - (1 - \mathbf{IC}(x^0))\lambda + B\lambda\varepsilon + C\lambda^2 < 0,$$

for some constants $A, B, C, D > 0$. Conditions (C_1') and (C_2') are polynomial constraints that define an admissible set for λ given ε and T . Standard algebraic manipulations show that this admissible set can be bounded in the form announced in Theorem 1.