Off-the-grid sparse regularization

Dongrui Shen

Abstract

This report presents an overview of Clarice Poon's seminar on *Off-the-grid sparse estimation* [4], focusing on the theoretical analysis of the stability of an off-the-grid sparse regularization method, Beurling-Lasso [6, 7]. First, we introduce the Beurling-Lasso method and some properties of the regularized problem which are critical to the stability analysis. Next we examine some of the related literature. Finally, in Section 3, we state the main result in an informal simplified way and point out the improvement compared with the related work.

1 Introduction

Sparse regularization is a technique used to prevent problems such as "overfitting" from occurring in machine learning and compressed sensing. Simply speaking, sparse regularization is just adding a "penalty" that favors sparser solution to the objective function. The regularized optimization problem could be expressed as

$$\min_{W,a} \frac{1}{2} L(W,a) + \|a\|_0, \tag{1.1}$$

where $\|\cdot\|_0$ is defined as the number of non-zero elements in a. In practice, we normally use a computationally feasible norm ℓ^1 to approximate the non-convex norm ℓ^0. In the case of linear regression, this method is known as Lasso. Further, we can generalize the Lasso method by considering a general problem of estimating an unknown sparse measure $\mu_0 \in \mathcal{M}(\mathcal{X})$, where $\mathcal{X} \subset \mathbb{R}^{d+1}$.

Problem 1 (Estimation of an unknown sparse measure)**.** *Define* $\phi(x) = (\varphi(\langle t_k, x^1\rangle + x^2))_k$, $\mathbf{m}(x) = \sum_{i=1}^n a_i \delta_{x_i} = \sum_{i=1}^n a_i \delta_{(w_i,b_i)}$, *where* $a_i \in \mathbb{R}$, $(w_i, b_i) \in \mathcal{X} \subset \mathbb{R}^{d+1}$. *Then the prediction model (multilayer perceptron) could be expressed as* $(f_{W,a}(t_k))_k = \Phi\mathbf{m} = \int_{\mathbb{R}^{d+1}} \phi(x)\mathrm{d}\mathbf{m}(x)$. *The corresponding convex optimization problem in the least squares sense is*

$$\min_{\mathbf{m}\in\mathcal{M}(\mathcal{X})} \|y - \Phi\mathbf{m}\|^2. \tag{1.2}$$

Beurling-Lasso (Blasso) is an increasingly popular method to estimate such a sparse measure. Its difference from Lasso is that the ℓ^1-norm is replaced by the "total variation".

Problem 2 (Beurling-Lasso)**.**

$$\min_{\mathbf{m}\in\mathcal{M}(\mathcal{X})} \frac{1}{2}\|y - \Phi\mathbf{m}\|^2 + \lambda\|\mathbf{m}\|_{\mathrm{TV}}, \tag{$\mathcal{P}_\lambda(y)$}$$

where the total variation $\|\mathbf{m}\|_{\mathrm{TV}} = \sup_{\{\mathcal{A}_i\}\subset\mathcal{X}} \sum_i |\mathbf{m}(\mathcal{A}_i)|$.

When $\mathbf{m}(x)$ is a sparse measure $\mathbf{m}_{a,x}$, $\|\mathbf{m}_{a,x}\|_{\mathrm{TV}} = \|a\|_1$. While when $\mathbf{m}(x)$ is continuous satisfying $\mathrm{d}\mathbf{m}(x) = f(x)\mathrm{d}x$, $\|\mathbf{m}_{a,x}\|_{\mathrm{TV}} = \|f\|_{L^1}$. Thus, we can regard it as a generalization of ℓ^1-norm from "the discrete setting" to "the continuous setting". In this study, we are interested in the theoretical properties of the regularized problem.

Proposition 1 (First order optimality condition)**.**

$$\mathbf{m}_\lambda(x) \in \mathrm{argmin}\, \mathcal{P}_\lambda(y) \Longleftrightarrow \eta_\lambda(x) = \frac{1}{\lambda}\Phi^*(y - \Phi\mathbf{m}_\lambda(x)) \in \partial\|\mathbf{m}_\lambda\|_{\mathrm{TV}}, \tag{1.3}$$

where Φ^* *denotes the corresponding pullback.*

In our simplified case, if $\mathbf{m}_\lambda$ is a sparse measure, where $\mathbf{m}_\lambda(x) = \sum_{i=1}^n a_i \delta_{x_i}$, then $\eta_\lambda \in \partial \|\mathbf{m}_\lambda\|_{\mathrm{TV}}$ means $\eta_\lambda \in \{\eta : \|\eta\|_\infty \leq 1, \eta(x_i) = \mathrm{sign}(a_i)\}$, where μ_λ is a solution to the dual problem of $(\mathcal{P}_\lambda(y))$, also known as "the dual certificate". It is the support of $\mathbf{m}_\lambda$ that we are really interested in. With the support, we can recover the sparse measure. In particular, it is contained in a set of "spikes" (or "amplitudes") of the dual certificate, i.e. $\mathrm{Supp}(\mathbf{m}_\lambda) \subset \{x : |\eta_\lambda(x)| = 1\}$ (see Fig. 1).

With the work of Duval and Peyré in 2015 [2], we only need to study the minimal norm certificate $\eta_0 = \lim_{\lambda \to 0} \eta_\lambda$ to understand the structure property of the recovered measure, where the limit is in the L^∞ sense. If η_0 is nondegenerate, then the sparsity and stability of Blasso solution from a noisy data is guaranteed.

2 Related work

From a theoretical perspective, understanding the performance of the Blasso method corresponds to establishing a "Rayleigh criterion", which is the minimum allowable separation distance between two spikes for the method to recover them from linear measurements. [6]

Candès and Fernandez-Granda [1] proved a sharp result under the Fourier measurements which is the first result in this direction. Tang et al [8] have extended this result to the case where only a small number of Fourier measurements are randomly selected. Notably, their result is only valid under a random signs assumption on the amplitudes of the Dirac masses and strongly depends on the translation invariance of the linear operator and the underlying domain, which are not applicable to spatially varying operators.

3 Main results

The main results obtained by Poon et al [6] are as follows. First, use the infinite-dimensional extension of the "golfing scheme" remove the random signs assumption of Tang et al while still keeping a sharp number of random measurements. Second, extend the framework to encompass non-translation invariant operators in a natural way with improved separation conditions.

The latter is done through a particular Riemannian geometric framework. To be precise, they first define the limit covariance kernel $K(x, x') = \mathbb{E}_\omega \overline{\varphi_\omega(x)} \varphi_\omega(x')$ which measures how much two Diracs at different points x, x' interact with each other as the number of samples approaches infinity. Then, they define the metric tensor in the Fisher metric sense, which is $\mathfrak{g}_x = \nabla_1 \nabla_2 K(x, x)$. Finally, define the associated geodesic distance,

$$d_{\mathfrak{g}}(x, x') = \inf_\gamma \int_0^1 \sqrt{\gamma'(t)^\top \mathfrak{g}_{\gamma(t)} \gamma'(t)} \mathrm{d}t, \qquad \text{(Fisher-Rao distance)}$$

where $\gamma : [0, 1] \to \mathcal{X}$, $\gamma(0) = x$, $\gamma(1) = x'$. The Fisher-Rao distance is the natural way to ensure and quantify support recovery, since it preserves the invariance of the problem under spatially varying operators [7]. Here we present an informal simplified version of their main result.

Theorem 1. *Let $s \in \mathbb{N}$ and let $(x_i)_{i=1}^s$ be s.t.*

$$\min_{i \neq j} d_{\mathfrak{g}}(x_i, x_j) \geq \Delta_{s,K}, \tag{3.1}$$

then η_0 is nondegenerate.

It shows that if the Fisher distance between spikes is larger than a Rayleigh separation constant, then the Blasso method recovers a stream of Diracs in a stable way. It is a general result in the multivariate setting, which is crucial for many practical applications. For example, in the quantitative magnetic resonance imaging problem (qMRI), fine discretization leads to highly coherent dictionaries which will break sparsitency. Formulation as sparse-group Blasso (SGB-Lasso) means we can have strong recovery guarantees (see Fig. 2 and Tab. 1) and the theoretical guarantees could be found in [5].

References

[1] E. J. Candès and C. Fernandez-Granda, *Towards a mathematical theory of super-resolution*, Communications on pure and applied Mathematics, 67 (2014), pp. 906–956.

[2] V. Duval and G. Peyré, *Exact support recovery for sparse spikes deconvolution*, Foundations of Computational Mathematics, 15 (2015), pp. 1315–1355.

[3] M. Golbabaee and C. Poon, *An off-the-grid approach to multi-compartment magnetic resonance fingerprinting*, arXiv preprint arXiv:2011.11193, (2020).

[4] C. Poon, *Off-the-grid sparse estimation.* ACM Seminar, University of Edinburgh, 2020.

[5] C. Poon and M. Golbabaee, *The sparse-group beurling-lasso*, tech. report, University of Bath, `https://cmhsp2.github.io/files/journal/sparse_group_blasso.pdf`, 2020.

[6] C. Poon, N. Keriven, and G. Peyré, *The geometry of off-the-grid compressed sensing*, arXiv preprint arXiv:1802.08464, (2018).

[7] C. Poon, N. Keriven, and G. Peyré, *Support localization and the fisher metric for off-the-grid sparse regularization*, in The 22nd International Conference on Artificial Intelligence and Statistics, PMLR, 2019, pp. 1341–1350.

[8] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, *Compressed sensing off the grid*, IEEE transactions on information theory, 59 (2013), pp. 7465–7490.

Appendix

The Fig. 1 below illustrates that the support of $\mathbf{m}_\lambda$ is contained in a set of "spikes" (or "amplitudes") of the dual certificate.

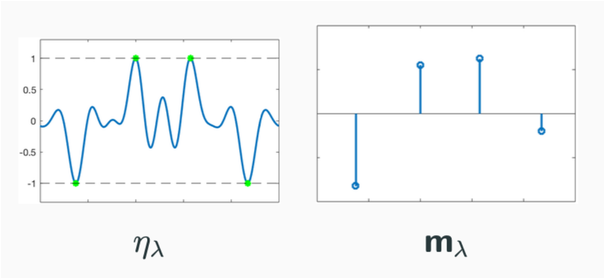

Figure 1: Characterization of regularized solutions $\mathbf{m}_\lambda$: $\mathrm{Supp}(\mathbf{m}_\lambda) \subset \{x : |\eta_\lambda(x)| = 1\}$.

Formulation as SGB-Lasso means we can have strong recovery guarantees. In Fig. 2, they compare the performance of the SGB-Lasso method with existing methods, SGB-Lasso outperforms the baselines in terms of the visual appearance of the mixture map. Further, the estimated T1/T2 values for WM and GM are within the range of literature values (see Tab. 1).

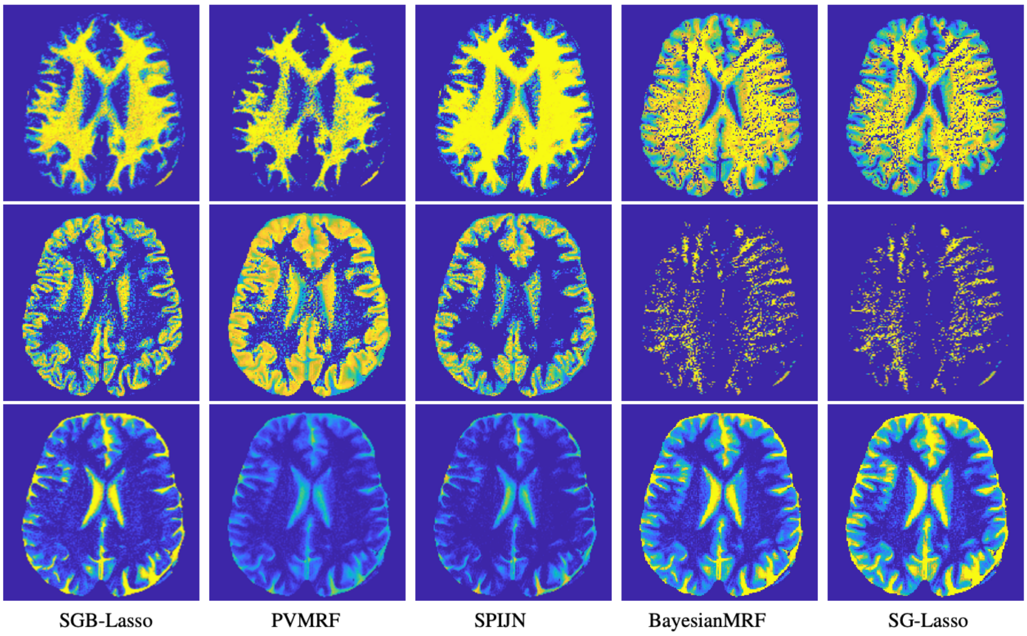

Figure 2: Estimated mixture maps of the WM, GM and a CSF related compartment for in-vivo brain using different methods [3].

	T1 (ms)					T2 (ms)				
Tissue	Literature	**SGB-Lasso**	PVMRF	SPIJN	BayesianMRF	Literature	**SGB-Lasso**	PVMRF	SPIJN	BayesianMRF
WM	694-862	829	806	699	821	68-87	81	80	51	77
GM	1074-1174	1114	1114	1483	874	87-103	102	105	164	82

Table 1: Estimated T1/T2 values for WM and GM compartments using different methods [3].