Fast Estimation of the Kernel Group LASSO

Zoltán Ádám Milacski*, Son Truong Hy, Balázs Pintér and András Lőrincz

Neural Information Processing Group, Department of Software Technology and Methodology, Faculty of Informatics, Eötvös Loránd University, Hungary

Abstract

The Kernel Group LASSO is an ℓ_1/ℓ_2 regularized (structured sparse) ℓ_2 reconstruction problem, which performs well at multi-label classification and is defined in a Reproducing Kernel Hilbert Space. Unfortunately, computing the ground truth solution to this task is slow for real-time applications even with state-of-the-art optimization schemes like the Fast Iterative Shrinkage Thresholding Algorithm. We extend the Learned Iterative Shrinkage Thresholding Algorithm – a fast neural network introduced by Gregor and LeCun – to estimate the true result. We test our method in time series classification by training on the 6D Motion Gesture Database while utilizing the Global Alignment time series kernel.

Kernel Group LASSO

Denote by $\boldsymbol{X} \neq \emptyset$ a set where the symmetric, positive semidefinite normalized kernel function $\boldsymbol{k}: \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ is defined:

$$\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{y}) = \frac{\langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{y}) \rangle_{\mathcal{H}}}{\sqrt{\langle \varphi(\boldsymbol{x}), \varphi(\boldsymbol{x}) \rangle_{\mathcal{H}}} \sqrt{\langle \varphi(\boldsymbol{y}), \varphi(\boldsymbol{y}) \rangle_{\mathcal{H}}}}, \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \boldsymbol{X} \times \boldsymbol{X} \tag{1}$$

for Reconstructing Kernel Hilbert Space $\mathcal{H}$ and feature mapping $\varphi: \boldsymbol{X} \to \mathcal{H}$. Let $\boldsymbol{x} \in \boldsymbol{X}$ and $\boldsymbol{D} = [\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N] \in \boldsymbol{X}$ be a signal and a vector system (i.e., a dictionary), respectively. Then for group structure $\mathcal{G} \subseteq 2^{\{1,\ldots,N\}}, \cup_{G \in \mathcal{G}} \boldsymbol{G} = \{1, \ldots, \boldsymbol{N}\}, \lambda > \boldsymbol{0}$, consider the **Kernel Group LASSO (KGLASSO)** procedure [1]:

$$\alpha^* = \operatorname*{arg\,min}_{\alpha \in \mathbb{R}^N} \frac{1}{2} \alpha^T \boldsymbol{k}(\boldsymbol{D}, \boldsymbol{D}) \alpha - \boldsymbol{k}(\boldsymbol{D}, \boldsymbol{x})^T \alpha + \lambda \sum_{G \in \mathcal{G}} \left(\sqrt{|\boldsymbol{G}|} \cdot \|\alpha_{\boldsymbol{G}}\|_2 \right), \tag{2}$$

i.e., we aim to **reconstruct $\varphi(\boldsymbol{x})$ with group sparse linear combination of $\varphi(\boldsymbol{D})$** within $\mathcal{H}$. This is an ℓ_1/ℓ_2 regularized quadratic programming problem that can be solved by the Fast Iterative Shrinkage Thresholding Algorithm (FISTA) [2] after precomputing $\boldsymbol{k}(\boldsymbol{D}, \boldsymbol{D})$ and $\boldsymbol{k}(\boldsymbol{D}, \boldsymbol{x})$.

Advantages:

- Kernels can discover *more subtle similarities* and make the system undercomplete.
- Group structure can *reduce the problem size* by choosing from fewer variables.
- Normalized group activations $\left(\frac{\|\alpha_G^*\|_2}{\sqrt{|G|}} \right)_{G \in \mathcal{G}}$ can be used for *multi-label classification*.

Limitations:

- FISTA still has significant *time complexity* as it requires several iterations.
- KGLASSO *scales quadratically in dictionary size* $\boldsymbol{N}$.
- KGLASSO *scales linearly in signal count.*
- Kernel computations can be very slow and often lead to *dense matrices*, which are difficult to store and deal with.

Learned Iterative Shrinkage Thresholding Algorithm

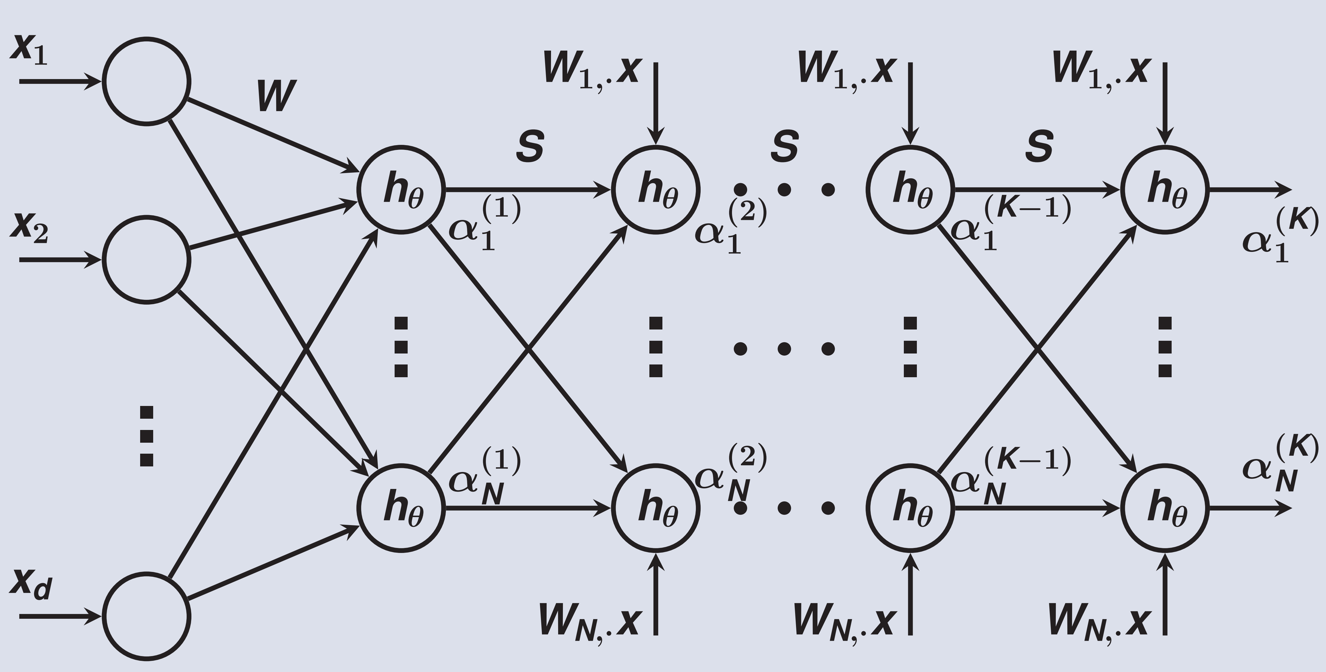
Due to the FISTA limitation above, supervised approximation schemes to it have gained attention recently. The **Learned Iterative Shrinkage Thresholding Algorithm (LISTA)** was proposed in [3]: a **neural network for estimating the sparse code** of the original linear and unstructured LASSO [4], where:

- *adaptive soft-thresholding activation function* was introduced to yield true sparse outputs with respect to tunable thresholds $\theta \in \mathbb{R}^N$:

$$h_\theta(\alpha) = \mathrm{sign}(\alpha) \circ (|\alpha| - \theta)_+, \tag{3}$$

- *competition between dictionary elements* was introduced by making the network recurrent with fixed depth $\boldsymbol{K} \in \mathbb{N}$:

$$\alpha^{(0)} = 0, \quad \alpha^{(k)} = h_\theta \left(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{S}\alpha^{(k-1)} \right), \quad k = 1, \ldots, K, \tag{4}$$



i.e., signal $\boldsymbol{x}$ is mapped to code space right away with $\boldsymbol{W}$ and then $\boldsymbol{S}$ rules out some of the active elements,

- *minimization of ℓ_2 loss* was carried out in parameters $\boldsymbol{W}, \boldsymbol{S}, \theta$ among a batch of training samples $(\boldsymbol{x}_{[i]}, \alpha^*_{[i]}), \boldsymbol{i} = 1, \ldots, \boldsymbol{M}$ with stochastic gradient descent and backpropagation through time:

$$L(\boldsymbol{W}, \boldsymbol{S}, \theta) = \frac{1}{2} \sum_{i=1}^{M} \|\alpha^*_{[i]} - \alpha^{(K)}_{[i]}\|_2^2. \tag{5}$$

Advantages:

- Matrix multiplications and soft-thresholding are *fast*.
- The algorithm has an *adjustable iteration count* $\boldsymbol{K}$.
- The method is *sparsity adaptive*, as θ is learnable.
- The *problem is reduced*, as only a correlated subset of all possible signals and sparse codes appear in an actual dataset.

Limitations:

- The scheme is limited to the case of the *linear and unstructured* LASSO problem.

Hypotheses

Our hypotheses:

- **LISTA may generalize to the structured case** via replacing outputs α^* with binary group activations $\left(\frac{\|\alpha_G^*\|_2}{\sqrt{|G|}} \right)_{G \in \mathcal{G}} > \boldsymbol{0}$, and computing a reduced pseudo-inverse.
- **LISTA may generalize to the kernelized case** via inputs $\boldsymbol{k}(\boldsymbol{D}, \boldsymbol{x})$ or $\left(\frac{\|\boldsymbol{k}(\boldsymbol{D}_G, \boldsymbol{x})\|_2}{\sqrt{|G|}} \right)_{G \in \mathcal{G}}$.
- **LISTA may bypass kernel computations** for further speedup via inputs $\boldsymbol{x}$ (as long as they are vectorizable).

To test our hypotheses, we implemented KGLASSO in Matlab and LISTA in Theano.

Experimental Setup

For our numerical experiment, we used *3D position features of uppercase air-handwriting characters* from the **6D Motion Gesture Database (6DMG)** [5]. The data contained **26** characters (A to Z) each repeated at most **10** times by **25** subjects. We normalized the data and uniformly interpolated each sample to length **128**. We then partitioned the set into train (**17** subjects) and test (**8** subjects) parts. Train samples were averaged after fixing both the subject and the character: the resulting **442** mean curves served as the dictionary $\boldsymbol{D}$. The test set supplied **102336** $\boldsymbol{x}_i$ signals by generating random convex combinations with additive noise after fixing both the subject and the character. Group structure $\mathcal{G}$ was induced by subjects ($\boldsymbol{G}$). $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{y})$ was chosen to be the **Global Alignment time series kernel** [6] with parameter $\sigma = \boldsymbol{0.9}$. We then set $\lambda = \boldsymbol{0.001}$ and computed ground truth binary group activations $\left(\frac{\|\alpha_G^*\|_2}{\sqrt{|G|}} \right)_{G \in \mathcal{G}} > \boldsymbol{0}$ according to equation (2). The average number of active groups thus became **5.2654**. The LISTA matrix $\boldsymbol{W}$ was pretrained with tied weights. For measuring **multi-label classification** performance, we applied an **80%-10%-10%** training-validation-testing shuffle-and-split scheme.

Results

Micro-averaged **multi-label classification results** were as follows.

Test set performance metric	$k(D,x)$		$\left(\frac{\|k(D_G,x)\|_2}{\sqrt{\|G\|}} \right)_{G \in \mathcal{G}}$		x	
	$K=1$	$K=2$	$K=1$	$K=2$	$K=1$	$K=2$
Loss function value	**0.300**	**0.329**	**1.497**	**1.077**	**1.323**	**1.052**
Accuracy	**0.962**	**0.961**	**0.793**	**0.850**	**0.758**	**0.820**
Precision	**0.930**	**0.924**	**0.824**	**0.781**	**0.599**	**0.682**
Recall	**0.950**	**0.954**	**0.418**	**0.712**	**0.671**	**0.789**
F_1 score	**0.940**	**0.939**	**0.555**	**0.745**	**0.633**	**0.732**

Conclusion

Our findings:

- **LISTA can generalize to the structured and kernelized KGLASSO case.**
- **Single layer is already capable.**
- **Mapping from kernels $\boldsymbol{k}(\boldsymbol{D}, \boldsymbol{x})$ is very accurate.**
- **Mapping from signals $\boldsymbol{x}$ the performance is considerable.**

Results may improve for larger $\boldsymbol{K}$ and with Convolutional neural network (CNN).

References

[1] László A Jeni, András Lőrincz, Zoltán Szabó, Jeffrey F Cohn, and Takeo Kanade. Spatio-temporal event classification using time-series kernel based structured sparsity. In *Computer Vision–ECCV 2014*, pages 135–150. Springer, 2014.

[2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[3] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proc. of the 27th Int. Conf. on Machine Learning (ICML-10)*, pages 399–406, 2010.

[4] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[5] Mingyu Chen, Ghassan AlRegib, and Biing-Hwang Juang. 6dmg: A new 6d motion gesture database. In *Proc. of the 3rd Multim. Sys. Conf.*, pages 83–88. ACM, 2012.

[6] Marco Cuturi. Fast global alignment kernels. In *Proc. of the 28th Int. Conf. on Machine Learning (ICML-11)*, pages 929–936, 2011.

This Activity is financially supported by EIT Digital.

 *Contact: srph25@gmail.com