TO BE OR NOT TO BE CONVEX? A STUDY ON REGULARIZATION IN HYPERSPECTRAL IMAGE CLASSIFICATION

Devis Tuia[1], *Remi Flamary*[2], *Michel Barlaud*[2]

[1] University of Zurich, Switzerland, devis.tuia@geo.uzh.ch
[2] Université de Nice-Sophia Antipolis, {remi.flamary,barlaud}@unice.fr

ABSTRACT

Hyperspectral image classification has long been dominated by convex models, which provide accurate decision functions exploiting all the features in the input space. However, the need for high geometrical details, which are often satisfied by using spatial filters, and the need for compact models (i.e. relying on models issued form reduced input spaces) has pushed research to study alternatives such as *sparsity inducing* regularization, which promotes models using only a subset of the input features. Although successful in reducing the number of active inputs, these models can be biased and sometimes offer sparsity at the cost of reduced accuracy. In this paper, we study the possibility of using *non-convex* regularization, which limits the bias induced by the regularization. We present and compare four regularizers, and then apply them to hyperspectral classification with different cost functions.

1. INTRODUCTION

Hyperspectral images (HSI) classification is one of the fast moving areas of modern remote sensing [1]. HSI poses new challenges for remote sensing image classification, in particular by their increased dimensionality and complexity. When confronted to HSI, parametric methods based on the estimate of the covariance matrix become either unfeasible or unreliable, since good estimations the class-covariance matrices require many labeled samples, which are usually not available. For these reasons, great research efforts are deployed in recent research to develop regularized approaches. Regularization allows to limit the model complexity in ill-posed situations characterized by high-dimensional data and limited number of samples and has proven very successful in both parametric [2] and non parametric [1] classification settings.

Most regularized approaches exploit the ℓ_2 norm, i.e. the squared sum of the model weights. The ℓ_2 norm has the advantage of being convex and many tools are available for convex optimization. However, squared norms provide active (non-zero) weights for all features, leading to models that are not compact. This can be a problem in HSI classification, wherethe use of spatial filters [3] increases the data dimensionality (and thus the number of coefficients to be estimated) considerably. For this reason, research in HSI classification and unmixing is turning to other types of regularization that induce sparsity [4, 5], i.e., the search for models where only a part of the initial coefficients is active. A deeply studied sparse regularizer is the ℓ_1 norm, or Lasso. Despite its desirable sparse nature, this norm is known to promote biased estimators and provides sparsity at the price of classification performance when strong sparsity is required. To obtain the highest performance, it is a common practice to run the ℓ_2-regularized classifier with the features selected by the sparse ℓ_1 model [6], but this is again against computational efficiency. Recently, nonconvex norms such as the ℓ_p-norm with $0 < p < 1$ have been proposed to limit the bias and obtain sparse and accurate predictions [7].

In this paper we provide a comparative study on different regularization strategies and study the joint behavior of performance and sparsity of each regularizer. We study four regularization strategies and deploy them with linear classifiers. The classifiers considered use a hinge and a calibrated loss function, respectively, and are applied on two challenging HSI classification tasks in urban and rural areas.

2. A FAMILY OF REGULARIZERS AND LOSS FUNCTIONS FOR HSI CLASSIFICATION

In this section we present a general optimization framework for HSI classification based on the estimation of linear prediction functions. Linear functions have regained popularity in numerous recent works, as they have shown state of the art performances when an adapted non-linear filtering is applied on the images. In other words, if the input space is well-selected and discriminative, a linear classifier can perform as accurately as a nonlinear classifier trained on the original input space. In the following, we therefore assume that we have the discriminative input space and focus on linear classifica-

This work was partly funded by the Swiss National Science Foundation under the grant PP00P2-150593.

tion only. Note that the performances shown in the experimental part are almost equivalent to the state of the art published in recent literature, thus confirming the hypothesis presented above.

2.1. Optimization problem

Let $\{\mathbf{x}_i, y_i\}_{i,\ldots,n}$ a set of labeled training samples where $y_i \in \{-1, 1\}$ is a binary class to be predicted and $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector that contains spectral bands, spatial filters or any kind of descriptor of the original pixels. We want to learn a linear prediction function of the form $f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b$ where $\mathbf{w} \in \mathbb{R}^d$ is the normal vector to the separating hyperplane and b is a bias term. The estimation is performed by solving the following regularized optimization problem:

$$\min_{\mathbf{w},b} \quad \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(\mathbf{x}_i)) + \lambda \sum_{j=1}^{d} g(|w_j|) \qquad (1)$$

where $\mathcal{L}(y, f(\mathbf{x}))$ is a data fitting term that measures the discrepancy between the prediction and the true label. The second term $g(\cdot)$ is a monotone function that defines the regularization term. λ weights the strenght of the regularization.

2.2. Data fitting term

The most commonly used data fitting term is the hinge loss defined as

$$\mathcal{L}_h(y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))^q,$$

When $q = 1$ this term boils down to the classical hinge loss used in Support Vector Machines (SVM) [8]. When $q = 2$ the loss is differentiable and thus easier to optimize with a gradient descent scheme [9]. In this work we will focus on the squared hinge loss with $q = 2$, which was already successfully applied on HSI classification problems [6].

Besides the traditional hinge loss, we will also investigate the use of the calibrated hinge loss defined in [10] as

$$\mathcal{L}_c(y, f(\mathbf{x})) = \max\{0, yf(\mathbf{x})\} - \ln(2 + |f(\mathbf{x})|),$$

This data fitting term is differentiable and is a strictly decreasing function. For these reasons, it can provide a posterior probability estimate as discussed in more detail in [10].

Regularization term	$g(\lvert w_j\rvert)$
Ridge, ℓ_2 norm	$\lvert w_j\rvert^2$
Lasso, ℓ_1 norm	$\lvert w_j\rvert$
Log sum penalty	$\log(\lvert w_j\rvert/\theta + 1)$
ℓ_p with $0 < p < 1$	$\lvert w_j\rvert^p$

Table 1. Definition of the regularization terms considered

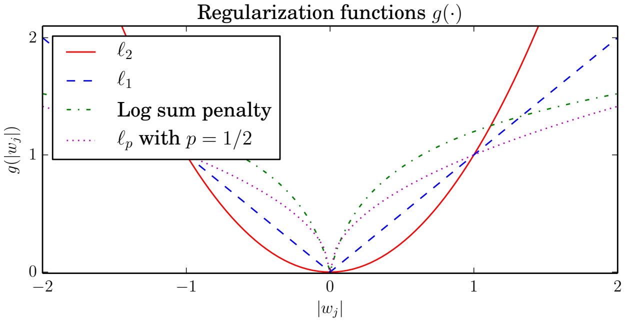

Fig. 1. Illustration of the regularization terms $g(\cdot)$. Note that both ℓ_2 and ℓ_1 regularizations are convex and that log sum penalty and ℓ_p with $p = 1/2$ are concave on their positive orthant.

2.3. Convex and non-convex regularization

The type of regularization is defined by the function $g(\cdot)$ in Eq. (1). The most common choice for this function is to use the square function, which leads to a regularization using the square of the Euclidean norm (ℓ_2) also known as ridge regularization. This regularization is commonly used for SVM [8] or for ridge regression. Another choice for $g(\cdot)$ is the identity function leading to a regularization term that consists in the sum of the absolute values of $\mathbf{w}$. This type of regularization, also known as Lasso [11] or ℓ_1, is non differentiable in 0 and promotes some components in $\mathbf{w}$ to be exactly 0. For this reason, ℓ_1 norms are often used for automatic feature selection during the optimization [6].

Despite its wide use, the Lasso is known to bias the estimators for sparse models [7, 12]. This bias is of particular importance in classification since it corresponds to a rotation on the hyperplane $\mathbf{w}$. To overcome this problem, authors in [6] trained *a posteriori* an ℓ_2 regularized classifier on the variables selected by the Lasso and obtained important gains in performances. Recently, the use of non-convex regularization for sparsity inducing regularization has been investigated as an alternative to overcome the bias of the Lasso without retraining of the classifier [12]. Among the most interesting, one can cite the log sum penalty regularizer proposed in [12] and the ℓ_p regularization with $p < 1$ [7]. All those regularizations are nonsmooth in 0 in order to promote sparsity but will impact less the large values of $|w_i|$ to limit the bias.

A definition of all the regularization terms investigated in this work is available in Table 1, along with an illustration of the amount of regularization in Fig. 1.

2.4. Optimization with non-convex regularization

Both the non-convex regularization terms above are non-differentiable. For this reason alternative optimization strategies with respect to convex optimization must be employed.

Among the algorithms for optimization with a non-differentiable regularization term are the Proximal Splitting methods [13]. The convergence of those algorithms to a global minimum are well studied in the convex case. For non-convex regularization, recent works have proved that proximal methods can be used with non-convex regularizers when a simple closed form solution of the proximity operator for the regularization can be computed [14]. In our case, we used the General Iterative Shrinkage and Thresholding (GIST) Algorithm proposed in [14]. We can use this algorithm, since both the data fitting terms $\mathcal{L}_h$ and $\mathcal{L}_c$ are gradient Lipschitz and all the regularization terms described above have a closed form proximal operator.

2.5. Extension to multiclass classification

In most HSI classification problems, there are K classes to be discriminated. In this case, we can readily adapt the optimization procedure discussed earlier by using a One-Against-All procedure. Such procedure consists of learning one linear function $f_k(\cdot)$ per class k and then predict the final class for a given observed pixel $\mathbf{x}$ as the solution of $\arg\min_k f_k(\mathbf{x})$. In practice, this leads to an optimization problem similar to (1), where a matrix $\mathbf{W}$, containing one classifier per class, needs to be estimated. The number of coefficients to be estimated is therefore the size d of the input space , multiplied by the number of classes.

3. DATA

In the experiments we consider two hyperspectral scenes: first is an image acquired in 2010 by the ProSpecTIR airborne system over the Pines sites in Indiana [15]. The image is composed by 445×750 pixels at $2m$ spatial resolution, with 360 spectral bands of $5nm$ width. Labels of sixteen types of agricultural landuse were available by a field survey (198‘074 labeled pixels). The second image is a CASI image acquired over Houston with 144 spectral bands at $2.5m$ resolution. The image depicts 14 urban land use classes, for which a field survey is also available (14‘703 labeled pixels). Additionally, a LiDAR DSM was also available and was used as an additional feature[1]. The CASI image was corrected with histogram matching for a large shadowed part on the right side and the DSM was detrended by a 3m trend on the left-right direction. For both datasets, we added contextual features to the spectral bands, in order to improve the geometric quality of classification [3]: we added morphological filters and texture filters, following the list in [6]. The filters were calculated using the 3 first principal components and with local sizes in the range $\{3, ..., 15\}$ pixels. The joint spatial-spectral input space is of dimension 540 in the Indiana image and 384 in the Houston data.

4. SETUP, RESULTS AND DISCUSSION

In the experiments below, we compare the four regularizers (ℓ_1, ℓ_2, Log sum penalty and ℓ_p with $p = 1/2$) and study the joint behavior of accuracy and sparsity along the regularization path, i.e. for different values of λ. To this end, we compute the solution of the optimization problem for regularization values $\lambda = \{1e^{-5}, \ldots, 1e^{-1}\}$, with 18 steps. For each step, the experiment was repeated five times with different train/test sets (each run with the same training samples for all regularizers and losses) and the average Kappa and number of active coefficients is reported in Fig 2. Please note that we report the total number of coefficients in the multiclass case, $w_{j,k}$, which is equal to the number of features multiplied by the number of classes, plus one additional feature per class (bias term). In the case of the Pines 2010 data, the model estimates 8‘656 coefficients, while in the Houston data, it deals with 5‘775. We repeat the same experiment for the two loss functions $\mathcal{L}_h$ and $\mathcal{L}_c$ with the linear classifier of Eq. (1). All the models are trained with 60 labeled pixels per class, randomly selected, and all the remaining labeled pixels are considered as the test set. Each test point in a 7×7 located in a spatial window around the training samples of the specific run are not taken into consideration.

The results are illustrated in Fig. 2, where the most desirable situation would be a classifier with both high accuracy and little active features, i.e., a score close to the top-left corner of the graphs. In all four settings, the ℓ_2 model (green dot) is remarkably accurate, but has all the coefficients ac-

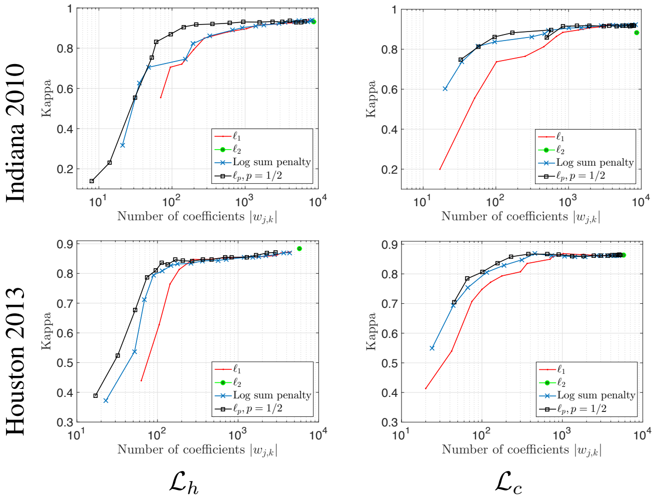

Fig. 2. Performance (Kappa) vs. compactness (number of coefficients $w_{i,k} > 0$) for the different loss functions and regularizers in the Pines 2010 and Houston datasets.

[1] The data were proposed as the Data Fusion Contest 2013 [16] and are available on the IADF TC website `http://www.grss-ieee.org/community/technical-committees/data-fusion/`

tive. Therefore, it is the less compact model. Employing the ℓ_1 regularizer (red line), as it is mainly done in the literature, achieves a sharp decrease in the number of active coefficients, but at the price of a steep decrease in performances of the classifier. When using 100 active coefficients, the ℓ_1 model suffers of a 20% drop in performance, a trend is observed in all the experiments reported.

Using the non-convex regularizers permits to have the best of both worlds: the ℓ_p regularizer (black line with '□' markers) and (in 3 out of 4 experiments) the Log sum penalty regularizer (blue line with '×' markers) achieve improvements of about 15-20% with respect to the ℓ_1 model with using 100 coefficients and show more stable results along the regularization path: the non-convex regularizers are indeed less biased than the ℓ_1 norm in classification and can achieve competitive performances with respect to the (non-sparse) ℓ_2 model with a fraction of the features (around 1-2%). If compact models are required, they seem thus to be a valid alternative to the classic non sparse ℓ_2 norm classifiers.

5. CONCLUSION

Sparsity is a characteristic of the greatest importance for future classification algorithms. The images have more bands and there is a great number of features that can be deployed. Being capable of selecting the important features only and to provide a classifier that is unbiased seems to be the way to go.

With these objectives in mind, we compared and studied four forms of regularization for linear classifiers and shown that the more recent non-convex regularization marry the accuracy of the classifier trained on a discriminative input space and the aim for compactness.

6. REFERENCES

[1] G. Camps-Valls, D. Tuia, L. Bruzzone, and J.A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Proc. Mag.*, vol. 31, no. 1, pp. 45–54, 2014.

[2] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Classification of hyperspectral images with regularized linear discriminant analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 3, pp. 862–873, 2009.

[3] M. Fauvel, Y. Tarabalka, J. A. Benediktsson, J. Chanussot, and J. C. Tilton, "Advances in spectral-spatial classification of hyperspectral images," *Proc. IEEE*, vol. 101, no. 3, pp. 652 – 675, 2013.

[4] M.-D. Iordache, J. M. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, 2011.

[5] Y. Chen, N.M. Nasrabadi, and T.D. Tran, "Hyperspectral image classification via kernel sparse representation," *IEEE Trans. Geosci. Rem. Sens.*, vol. 51, no. 1, pp. 217–231, 2013.

[6] D. Tuia, M. Volpi, M. Dalla Mura, A. Rakotomamonjy, and R. Flamary, "Automatic feature learning for spatio-spectral image classification with sparse SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6062–6074, 2014.

[7] T. Zhang, "Analysis of multi-stage convex relaxation for sparse regularization," *J. Mach. Learn. Res.*, vol. 11, pp. 1081–1107, 2010.

[8] B. Schölkopf, C. J.C. Burges, and A. J. Smola, *Advances in kernel methods: support vector learning*, MIT press, 1998.

[9] O. Chapelle, "Training a support vector machine in the primal," *Neural Comput.*, vol. 19, no. 5, pp. 1155–1178, 2007.

[10] W. Bel Haj Ali, R. Nock, and M. Barlaud, "Boosting stochastic newton with entropy constraint for large-scale image classification," in *Proc. ICPR*, Stockholm, Sweden, 2014.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc. B*, pp. 267–288, 1996.

[12] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted 1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[13] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*, pp. 185–212. Springer, 2011.

[14] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye, "A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems," in *Proc. ICML*, Atlanta, GE, 2013.

[15] M. M. Crawford, D. Tuia, and L. H. Hyang, "Active learning: Any value for classification of remotely sensed data?," *Proceedings of the IEEE*, vol. 101, no. 3, pp. 593–608, 2013.

[16] F. Pacifici, Q. Du, and S. Prasad, "Report on the 2013 IEEE GRSS data fusion contest: Fusion of hyperspectral and LiDAR data," *IEEE Remote Sens. Mag.*, vol. 1, no. 3, pp. 36–38, 2013.