Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Multi Class Learning with Individual Sparsity

Ben Zion Vatashsky and **Koby Crammer**
Department of Electrical Engineering
Technion - Israel Institute of Technology, 32000 Haifa, Israel
vatashsky@gmail.com, koby@ee.technion.ac.il

Abstract

Multi class problems are everywhere. Given an input the goal is to predict one of a few possible classes. Most previous work reduced learning to minimizing the empirical loss over some training set and an additional regularization term, prompting simple models or some other prior knowledge. Many learning regularizations promote sparsity, that is, small models or small number of features, as performed in group LASSO. Yet, such models do not always represent the classes well. In some problems, for each class, there is a small set of features that represents it well, yet the union of these sets is not small. We propose to use other regularizations that promote this type of sparsity, analyze the generalization property of such formulations, and show empirically that indeed, these regularizations not only perform well, but also promote such sparsity structure.

1 Introduction

Regularization is a highly-used and understood task in supervised learning. Given data-samples, called training data, modern algorithms not only seek for a model that performs well on the training data, but also require that the model would be simple in some sense, where simplicity is measured via some regularization function. Regularization is widely used as a mechanism to prevent overfitting or impose prior knowledge of a structure on a model. Learning binary prediction problems or regression of real numbers is studied for over half a century, with a lot of work focusing in linear models. Given an input $\boldsymbol{x} \in \Re^d$, its inner product with some vector $\boldsymbol{\omega} \in \Re^d$ is used to make a prediction, $f(\boldsymbol{\omega}^\top \boldsymbol{x})$. In many cases regularization is defined to be some norm of that vector, $\|\boldsymbol{\omega}\|$.

Among others, SVM [Boser *et al.*, 1992; Cortes and Vapnik, 1995] can be represented as such classification problem with ℓ_2 norm [Mukherjee *et al.*, 2002]. For regression, Ridge Regression [Hoerl, 1962; Hoerl and Kennard, 1970; Tikhonov and Arsenin, 1977] (ℓ_2 norm) and LASSO [Tibshirani, 1996] (ℓ_1 norm) are popular.

In more complex problems, such as multi class categorization, in which given an input $\boldsymbol{x}$ the algorithm is required to output one of c possible classes, matrix models $\boldsymbol{\omega} \in \Re^{d \times c}$ are used. Here, linear models prediction is performed by first computing the inner product of the input with each column of the matrix, and then processing the resulting c scalars, $f(\boldsymbol{\omega_1}^\top \boldsymbol{x}, ..., \boldsymbol{\omega_c}^\top \boldsymbol{x})$. In multi class problems we have $f(a_1, ..., a_c) = \arg\max_s a_s$. Natural extensions for norm regularization from vector to matrices are entry-wise norms, such as the Frobenius norm: $\sqrt{\sum_s \sum_t |\boldsymbol{\omega}_{t,s}|^2}$, used in multi class SVMs [Weston and Watkins, 1998; Crammer and Singer, 2001; Lee *et al.*, 2004].

Alternatively a mixed norm can be used. The most common such usage is group LASSO [Yuan and Lin, 2006; Bakin, 1999] which uses the $\ell_{2,1}$ mixed norm as a mechanism for selecting a group of variables. That is, the regularization promotes choosing a small sub-set of the features (rows), and then any feature from this subset can be used. However, if we seek sparser models, then the $\ell_{1,1}$ (which is the vector ℓ_1 norm in the mixed norm representation) is often used, as a convex relaxation of counting the number of non-zero elements.

We argue that for many situations these two norms are not capturing our requirements from models. Though it is reasonable to believe that many features are redundant, this redundancy might be different among classes. For example color pattern may be very informative for zebras, but less informative for horses, dogs, cats or chameleons which may have a variety of color patterns, and thus $\ell_{2,1}$ may not be the right choice. Additionally, the global sparsity prompted by $\ell_{1,1}$ may generate models that are very sparse for some classes, and dense for others. This may happen if the data is far from being balanced, as there are few examples of one class, and many of another classes.

In this work we propose to use another principle for regularization. Instead of forcing a small number of features ($\ell_{2,1}$) or a small model altogether ($\ell_{1,1}$), we propose to promote models, in which for each class (independently) there would be a small number of relevant features. Such regularization would generate small, that is sparse, models, yet would not "favor" certain classes. We formulate learning with such regularization in the next section and provide robustness analysis and generalization bound for this, and in fact, for all mixed norms. Specifically, we show that our regularization is equivalent to an algorithm which is robust to a feature noise that is

different per class, yet worst classes' noise is not too large.

We also report results with 14 text classification problems, with a large range in size, number of classes and dimensions. We show that our proposed regularization attains higher F_1 scores than any other mixed norm regularization we tried, yet results in sparser models. One explanation for that difference is that our algorithm attains higher recall values (over classes) paying in precision. This demonstrates exactly the point that the obtained models are not too-sparse for some classes, as models learned with $\ell_{1,1}$ regularization may be (when some classes contain very few non-zero model terms, they may never be predicted).

Related work: Most work with mixed norm regularization focussed on $\ell_{2,1}$ and $\ell_{\infty,1}$ norms [Duchi and Singer, 2009b; 2009a; Bradley and Bagnell, 2009; Zhao *et al.*, 2009; Mairal *et al.*, 2010]. A work on mixed norm regularization including $\ell_{1,2}$ was proposed for signal estimation [Kowalski, 2009; Kowalski and Torrésani, 2008] and kernel learning for binary classification [Kowalski *et al.*, 2009]. In a work on hierarchical penalization [Szafranski *et al.*, 2008] a $\ell_{\frac{4}{3},1}$ norm was derived and used. The $\ell_{p,1}$ norm regularization was also used in other contexts [Fornasier and Rauhut, 2008; Teschke and Ramlau, 2007].

Notation: Given a matrix A, its mixed norm, $\ell_{p,q;r}$, is defined by computing the p norm of each column (or row) of A, and then computing the q norm of the result. The order of the summation (columns or rows first) is defined by $r \in \{1,2\}$. We define a mixed norm where we either first sum over rows ($r=1$) or over columns ($r=2$),

$$\|A\|_{p,q;1} = \left(\sum_t \left(\sum_s |A_{t,s}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}}$$

$$\|A\|_{p,q;2} = \left(\sum_s \left(\sum_t |A_{t,s}|^p \right)^{\frac{q}{p}} \right)^{\frac{1}{q}} .$$

An early work on mixed norms spaces was done by Benedek and Panzone at 1961 [Benedek and Panzone, 1961].

2 Individual Variable Selection

Group selection, promoted by group LASSO and other mixed norms of the form $\ell_{p,1}$, assume that one group of variables have good descriptive qualities for the entire problem. Specifically, for feature selection in multi class, the assumption is that a subset of the features is globally able to describe well all the classes. That is, a small set is sought that would work well across all classes. We take a different route and assume that for each class there is a small number of features that describe it well. However, we do not assume that this set overlaps other classes' sets, and also, we do not want to have too small number of features for some classes. This intuition leads to the following regularization, $\sum_i (\text{zero norm of column i})^2$. That is, we compute the number of non-zero elements per class and then take the Euclidean norm of this vector of counts. The Euclidean norm is used, as we want to demote classes for which there are many non-zero elements compared with others. As performed in other contexts, we replace the zero norm with the unit norm, which is convex and continuous, and obtain the following relaxation,

$$\Omega(\boldsymbol{\omega}) = \frac{1}{2} \|\boldsymbol{\omega}\|_{1,2;2}^2 = \frac{1}{2} \sum_{s=1}^{c} \left(\sum_{t=1}^{d} |\boldsymbol{\omega}_{t,s}| \right)^2 .$$

Thus, given a training set $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i \in \Re^d$ and $y_i \in \{1,...,c\}$, we propose to learn by minimizing the following problem,

$$\min_{\boldsymbol{\omega}} \sum_{i=1}^{n} L(\boldsymbol{\omega}, (\boldsymbol{x}_i, y_i)) + \frac{\lambda}{2} \|\boldsymbol{\omega}\|_{1,2}^2 ,$$

where $L(\boldsymbol{\omega}, (\boldsymbol{x}_i, y_i))$ is some multi class loss, such as the multi class log loss (with a curvature parameter ν),

$$L_{\log}(\boldsymbol{\omega}, \boldsymbol{x}, y^*) = \frac{1}{\nu} \log \left(1 + \sum_{y \neq y^*} e^{\nu(1 + \boldsymbol{\omega}_y \boldsymbol{x} - \boldsymbol{\omega}_{y^*} \boldsymbol{x})} \right) . \quad (1)$$

Previous work with $\ell_{1,2}$ regularization included signal reconstruction [Kowalski, 2009; Kowalski and Torrésani, 2008] and binary classification with multiple kernels [Kowalski *et al.*, 2009]. To the best of our knowledge no work has used $\ell_{1,2}$ regularization for multi class problems.

3 Analysis

We now provide an analysis for learning with mixed-norm matrices. First we state an equivalence to robustness over a certain noise, and then a generalization bound based on Gaussian complexity analysis. In both cases we build on previous tools, and modify them to our setting. We emphasize that it is not a direct application, but a non-trivial derivation is needed.

The properties we show in this chapter demonstrate theoretical guarantees according to a prior knowledge. This prior knowledge may guide us to choose our proposed $\ell_{1,2;2}$ regularization. We extend the equivalence of binary SVM to robust optimization [Xu *et al.*, 2009] to a multi class setting. We also state generalization bounds for $\ell_{1,2;2}$ regularization, using Gaussian complexity.

3.1 Equivalence to noise robustness

Robustness to noise is one of regularization's main goals. Recently Xu et al. [Xu *et al.*, 2009] showed, that with some limitations, regularization for binary classification using hinge loss, is equivalent to robust optimization. We extend this notion to the multi class setting by showing an equivalence of mixed norm regularization and robustness to noise for the sum of hinges loss function.

One approach we tried was to reduce multi class learning into a binary classification problem with c-times more examples. Applying the result of Xu et al. [Xu *et al.*, 2009] on the new problem, provided us a robustness equivalence only to $\ell_{p,1;2}$ mixed norms.

We thus re-derived this equivalence from scratch. Following the outline of Xu et al. [Xu *et al.*, 2009] we proved the following result (here $\boldsymbol{y}_i \in \{-1,1\}^c$).

Theorem 1 *Given a set of examples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n$, non separable for each class $s = 1, ..., c$ (for each class $s = 1, ..., c$ and every $\boldsymbol{\omega}_s \in \Re^d$, there is an example $j(s)$ such that: $y_{j(s),s}(\boldsymbol{\omega}_s^\top \mathbf{x}_{j(s)}) < 0$, where $y_{j(s),s}$ is the s term of $\boldsymbol{y}_{j(s)}$), and the two sets,*

$$\check{\mathcal{N}} = \left\{ (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_{\boldsymbol{n}}) \;\middle|\; \sum_{i=1}^n \|\boldsymbol{\delta}_i\|_{p_*,q_*} \leq M \right\}$$

and

$$\check{\mathcal{N}}_0 = \left\{ \boldsymbol{\delta} \;\middle|\; \|\boldsymbol{\delta}\|_{p_*,q_*} \leq M \right\} ,$$

for some $p_*, q_* \geq 1$, the following two optimization problems are equivalent:

$$\min_{\boldsymbol{\omega}} \sup_{\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_n \in \check{\mathcal{N}}} \left\{ \sum_{i=1}^n \sum_{s=1}^c \left[1 - y_{i,s}(\boldsymbol{\omega}_s^\top (\mathbf{x}_i - \boldsymbol{\delta}_{i,s}))\right]_+ \right\}$$

$$\min_{\boldsymbol{\omega},\boldsymbol{\xi}} \sup_{\boldsymbol{\delta} \in \check{\mathcal{N}}_0} \langle \boldsymbol{\omega}, \boldsymbol{\delta} \rangle + \sum_{i=1}^n \sum_{s=1}^c \xi_{i,s}$$

$$\text{s.t.} \quad \xi_{i,s} \geq 1 - y_{i,s}(\boldsymbol{\omega}_s^\top \mathbf{x}_i) \qquad i = 1, ..., n;\ s = 1, ..., c$$
$$\xi_{i,s} \geq 0 \qquad i = 1, ..., n;\ s = 1, ..., c \quad , \tag{2}$$

where $\langle A, B \rangle = Tr(A^\top B)$ is the matrix (Frobenius) inner product.

The proof is omitted due to lack of space. When a norm of the perturbation is bounded, the regularization term is similar to the definition of the dual norm (multiplied by the bound value). This means that ℓ_{p_*,q_*} is the dual norm of some $\ell_{p,q}$ norm, where $\frac{1}{p} + \frac{1}{p_*} = \frac{1}{q} + \frac{1}{q_*} = 1$ [Bradley and Bagnell, 2009]. Thus, we get the regularization term:

$$\sup_{\|\boldsymbol{\delta}\|_{p_*,q_*} \leq M} \langle \boldsymbol{\omega}, \boldsymbol{\delta} \rangle = M \|\boldsymbol{\omega}\|_{p,q} . \tag{3}$$

According to Eq. (3), the correspondence between the regularization tradeoff parameter λ and the noise bound M is $\lambda = M$. In other words, λ may be used to estimate the noise bound M ($= \lambda$) and vice versa. For example, if we know that $\sum_{i=1}^n \|\boldsymbol{\delta}_i\|_{\infty,2;2} \leq M$, then $\ell_{1,2;2}$ would be the best regularization. In this type of noise, for each class the noise may be different, yet it does not get too high, as the norm of largest classes' noise, summed over examples should not be too large.

3.2 Generalization bound

We next provide a generalization analysis for the $\ell_{1,2;2}$ regularization using Gaussian complexity. We use the following measure of function complexity given by Bartlett and Mendelson [Bartlett and Mendelson, 2003]:

Definition 1 *[Bartlett and Mendelson, 2003] The Gaussian complexity of a function class $\mathcal{F}$ mapping from a set $\mathcal{X}$ to $\Re$ is defined as:*

$$\mathcal{G}_n(\mathcal{F}) = \mathbb{E}_X \mathbb{E}_g \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_1^n g_i f(X_i) \right| \right] ,$$

where $X_1, ..., X_n$ are samples selected independently from the set $\mathcal{X}$ according to a probability $\mathcal{P}$ and $g_1, ..., g_n$ are independent Gaussian random variables, where for each $i \in \{1, ..., n\}$, $g_i \sim \mathcal{N}(0, 1)$.

Based on this definition Bartlett and Mendelson proved the following theorem.

Theorem 2 *[Bartlett and Mendelson, 2003] Let $\mathcal{F}$ be a class of function mapping from $\mathcal{X}$ to $\mathcal{A} = \Re^c$ and let $\mathcal{F}_1, ...\mathcal{F}_c$ be real valued classes, such that $\mathcal{F}$ is a subset of their direct sum. For a given loss function $L\colon \mathcal{Y} \times \mathcal{A} \to [0, 1]$, let $\phi\colon \mathcal{Y} \times \mathcal{A} \to [0, 1]$ be a dominating cost function (for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, $\phi(y, a) \geq L(y, a)$), such that for all $y \in \mathcal{Y}$, $\phi(y, \cdot)$ is a Lipschitz function (with respect to Euclidean distance on $\mathcal{A}$) with a constant Λ. Let $(X_i, Y_i)_{i=1}^n$ be samples selected independently according to probability $\mathcal{P}$. Then, for any integer n and $0 < \delta < 1$, with a probability of at least $1 - \delta$, over samples of size n, for every $f \in \mathcal{F}$, the following holds:*

$$\mathbb{E}L(Y, f(X)) \leq \hat{\mathbb{E}}_n \phi(Y, f(X)) + k\Lambda \sum_{s=1}^c \mathcal{G}_n(\mathcal{F}_s) + \sqrt{\frac{8 \ln(2/\delta)}{n}} , \tag{4}$$

where k is some constant.

We focus on the sum of clipped hinges loss,

$$\phi(Y, \boldsymbol{f}(x)) = \frac{1}{c} \sum_{s=1}^c \min\left\{1, \max\left\{0, 1 - Y_s \cdot \left(\boldsymbol{\omega}_s^\top X\right)\right\}\right\},$$

$$\hat{\mathbb{E}}_n \phi(Y, \boldsymbol{f}(X)) = \frac{1}{n} \sum_{i=1}^n \phi(Y_i, \boldsymbol{f}(X_i)) \text{ and } f_s \in \mathcal{F}_s^M$$

such that:

$$\mathcal{F}_s^M = \Big\{ \boldsymbol{x} \mapsto \boldsymbol{\omega}_s^\top \boldsymbol{x} \colon \boldsymbol{\omega}_s \in \Re^d,\ s \in \{1, ..., c\},$$
$$\boldsymbol{\omega} = [\boldsymbol{\omega}_1 ... \boldsymbol{\omega}_c],\ \|\boldsymbol{\omega}\|_{p,q;r} \leq M \Big\} \tag{5}$$

and $\boldsymbol{f} = (f_1, ..., f_c) \in \mathcal{F}^M = \left(\mathcal{F}_1^M, ..., \mathcal{F}_c^M\right)$.

In the next lemma we bound the Gaussian complexity term in Eq. (4) for our setting.

Lemma 3 *Let $\boldsymbol{f} = \{f_1, ..., f_c\} \in \mathcal{F}^M = \left(\mathcal{F}_1^M, ..., \mathcal{F}_c^M\right)$, belong to a set defined in Eq. (5). Then, the following bound holds for the corresponding Gaussian complexities $\mathcal{G}_n\left(\mathcal{F}_s^M\right)$:*

$$\sum_{s=1}^c \mathcal{G}_n\left(\mathcal{F}_s^M\right) \leq \frac{2M}{n} \mathbb{E}_X \mathbb{E}_g \|\mathbf{X}\boldsymbol{g}\|_{p_*,q_*;r} \quad ,$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \Re^{d \times n}$ is a matrix of n samples, independently selected, according to probability $\mathcal{P}$ and $\boldsymbol{g} \in \Re^{n \times c}$ is a matrix with independent Gaussian variables $g_{s,i}$.

The proof is omitted due to lack of space.

We now state and prove the main theorem of the section.

Theorem 4 *Let $\boldsymbol{f} = (f_1, ..., f_c) \in \mathcal{F}^M = \left(\mathcal{F}_1^M, ..., \mathcal{F}_c^M\right)$, defined in Eq. (5) with a bounded $\ell_{1,2;2}$ norm. Then, for any integer n and $0 < \delta < 1$, with a probability of at least $1 - \delta$, over samples of length n, the following holds:*

$$\mathbb{E}L(Y, f(X)) \leq \hat{\mathbb{E}}_n \phi(Y, \boldsymbol{f}(X)) + 2kM\sqrt{\frac{2c \ln 2d}{n}} X_\infty^{UB} + \sqrt{\frac{8 \ln(2/\delta)}{n}} ,$$

where X_∞^{UB} is an upper bound on the ℓ_∞ norm of $\mathbf{x}$.

Proof: Plugging Lemma 3 in Theorem 2, and using the fact that the Lipschitz constant for the sum of hinges loss is 1, and that $\mathbb{E}_X[\cdot] \leq \sup_{X\in\mathcal{X}}[\cdot]$, we get the following generalization bound:

$$\mathbb{E}L(Y,f(X)) \leq \hat{\mathbb{E}}_n\phi(Y,\boldsymbol{f}(X)) + \frac{2kM}{n}\sup_{X\in\mathcal{X}}\mathbb{E}_g\left\|\mathbf{X}\boldsymbol{g}\right\|_{p_*,q_*;r} + \sqrt{\frac{8\ln(2/\delta)}{n}}. \tag{6}$$

For $p=1$, $q=2$, $r=2$ we have $p_*=\infty$, $q_*=2$, $r_*=2$,

$$\sup_{X\in\mathcal{X}}\mathbb{E}_g\left\|\mathbf{X}\boldsymbol{g}\right\|_{\infty,2;2} \leq \sup_{X\in\mathcal{X}}\mathbb{E}_g\left\|\mathbf{X}\boldsymbol{g}\right\|_{\infty,1;2} = \sup_{\boldsymbol{x}_i\in\mathcal{X}}\mathbb{E}_g\sum_{s=1}^{c}\max_{t=1\ldots d}\left|\sum_{i=1}^{n}x_{i,t}g_{i,s}\right|. \tag{7}$$

Next we bound the right expectation, using a technique used by Massart [Massart, 2000]. Using Jensen's inequality, we get for all $\mu>0$:

$$\begin{aligned}
&\exp\left(\mu\mathbb{E}_g\left[\sum_{s=1}^{c}\max_{t=1\ldots d}\left|\sum_{i=1}^{n}x_{i,t}g_{i,s}\right|\right]\right)\\
\leq&\mathbb{E}_g\left[\exp\left(\mu\sum_{s=1}^{c}\max_{t=1\ldots d}\left|\sum_{i=1}^{n}x_{i,t}g_{i,s}\right|\right)\right]\\
=&\mathbb{E}_g\left[\prod_{s=1}^{c}\max_{t=1\ldots d}\exp\left(\mu\left|\sum_{i=1}^{n}x_{i,t}g_{i,s}\right|\right)\right]\\
\leq&\prod_{s=1}^{c}\sum_{t=1}^{d}\mathbb{E}_g\left[\exp\left(\mu\left|\sum_{i=1}^{n}x_{i,t}g_{i,s}\right|\right)\right]\\
\leq&2\prod_{s=1}^{c}\sum_{t=1}^{d}\mathbb{E}_g\left[\exp\left(\mu\sum_{i=1}^{n}x_{i,t}g_{i,s}\right)\right]\\
=&2\prod_{s=1}^{c}\sum_{t=1}^{d}\prod_{i=1}^{n}\mathbb{E}_g\left[\exp\left(\mu x_{i,t}g_{i,s}\right)\right]\\
=&\frac{2}{\sqrt{2\pi}}\prod_{s=1}^{c}\sum_{t=1}^{d}\prod_{i=1}^{n}\int_{-\infty}^{\infty}\exp\left(\mu x_{i,t}g_{i,s}-\frac{1}{2}g_{i,s}^2\right)dg_{i,s}.
\end{aligned}$$

Using the Gaussian integral, $\int_{-\infty}^{\infty}e^{-x^2}dx=\sqrt{\pi}$, we bound the RHS of the last equality with,

$$\begin{aligned}
2\prod_{t=s}^{c}\sum_{t=1}^{d}\prod_{i=1}^{n}\exp\left(\frac{1}{2}\mu^2x_{i,t}^2\right)=&2\prod_{s=1}^{c}\sum_{t=1}^{d}\exp\left(\frac{1}{2}\mu^2\sum_{i=1}^{n}x_{i,t}^2\right)\\
\leq&2d\exp\left(\frac{1}{2}\mu^2\sum_{s=1}^{c}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|_\infty^2\right).
\end{aligned}$$

By taking log and dividing by μ, we get (for $\mu>0$):

$$\mathbb{E}_g\left[\sum_{s=1}^{c}\max_{t=1\ldots d}\left|\sum_{i=1}^{n}x_{i,t}g_{i,s}\right|\right] \leq \frac{\ln 2d}{\mu}+\frac{1}{2}\mu\sum_{s=1}^{c}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|_\infty^2.$$

Setting $\mu=\sqrt{\frac{2\ln 2d}{\sum_{s=1}^{c}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|_\infty^2}}$, leads to the following result,

$$\begin{aligned}
\mathbb{E}_g\left[\sum_{s=1}^{c}\max_{t=1\ldots d}\left|\sum_{i=1}^{n}x_{i,t}g_{i,s}\right|\right] \leq&\sqrt{2\sum_{s=1}^{c}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|_\infty^2\ln 2d}\\
=&\sqrt{2c\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|_\infty^2\ln 2d}. \qquad (8)
\end{aligned}$$

We denote by X_∞^{UB} any bound on the ℓ_∞ norm of $\boldsymbol{x}$. Substituting Eq. (8) in Eq. (7) yields,

$$\sup_{X\in\mathcal{X}}\mathbb{E}_g\left\|\mathbf{X}\boldsymbol{g}\right\|_{\infty,2;2} \leq X_\infty^{\text{UB}}\sqrt{2nc\ln 2d}. \tag{9}$$

Plugging the right term of Eq. (9) in Eq. (6) would produce the desired bound. ■

Using other norms would yield different terms in Eq. (9). This demonstrates that the norm may be selected according to bounds and parameters of the data.

4 Multi Class Learning Optimization

As mentioned above, our multi class optimization problem is a sum of a loss term and a regularization mixed norm term:

$$\min_{\boldsymbol{\omega}}\mathfrak{L}(\boldsymbol{\omega},\{(\boldsymbol{x}_i,y_i)\})+\frac{\lambda}{q}\left\|\boldsymbol{\omega}\right\|_{p,q}^q, \tag{10}$$

where $\mathfrak{L}(\boldsymbol{\omega},\{(\boldsymbol{x}_i,y_i)\}) = \sum_{i=1}^{n}L(\boldsymbol{\omega},\{(\boldsymbol{x}_i,y_i)\})$ is the losses sum.

We implemented a proximal splitting algorithm [Combettes and Pesquet, 2011] to solve this problem which is convex, and is composed by a sum of a smooth function and a non-differentiable function. The algorithm we used is the forward-backward splitting which is suitable for such problems.

The forward-backward algorithm is iterating between two steps: a gradient step for the smooth function - called forward step, and a proximity step - called backward step, where the proximity operator of the second function is used on the result of the previous step. A pseudo code of the algorithm is given in Alg. 1.

For the gradient step we use an adaptive learning rate γ_k for each iteration k. The value of γ_k is selected by starting with high value and reducing it until the following condition is met [Bach *et al.*, 2011],

$$\mathfrak{L}(\boldsymbol{\omega}_{new}) \leq \mathfrak{L}(\boldsymbol{\omega}_k)+\text{Tr}\left(\nabla\mathfrak{L}(\boldsymbol{\omega}_k)^\top(\boldsymbol{\omega}_{new}-\boldsymbol{\omega}_k)\right)+\frac{1}{2\gamma_k}\left\|\boldsymbol{\omega}_{new}-\boldsymbol{\omega}_k\right\|_2^2. \tag{11}$$

In our experiments we found that indeed this method produced better results than with a constant γ. For the third step of the algorithm we use the proximity operator defined by Moreau [Moreau, 1962] for lower semicontinuous convex functions, adapted here for mixed norms,

$$\text{prox}_{\frac{\gamma\lambda}{q}\|\cdot\|_{p,q}^q}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\omega}}\frac{\gamma\lambda}{q}\left\|\boldsymbol{\omega}\right\|_{p,q}^q+\frac{1}{2}\left\|\boldsymbol{\omega}-\boldsymbol{\theta}\right\|_2^2.$$

Algorithm 1 Forward-backward algorithm for multi class with $\ell_{p,q}$ regularization

Initialize: $\boldsymbol{\omega}_0$, $k = 0$
Iterate:

1. set γ_k according to 11
2. $\boldsymbol{\theta}_k = \boldsymbol{\omega}_k - \gamma_k \nabla_{\boldsymbol{\omega}_k} \mathfrak{L}(\boldsymbol{\omega}_k, \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\})$
3. $\boldsymbol{\omega}_{k+1} = \text{prox}_{\frac{\gamma_k \lambda}{q} \|\cdot\|_{p,q}^q}(\boldsymbol{\theta}_k)$
4. $k \leftarrow k + 1$

Proximity operators for mixed norms were developed by Kowalski [Kowalski, 2009], with closed expressions for $p, q \in \{1, 2\}$. These operators were used for binary kernel classification [Kowalski *et al.*, 2009].

5 Empirical study

In order to empirically analyze $\ell_{1,2;2}$ regularization, we evaluated all six regularization functions of the combinations $p, q \in \{1, 2\}$, combined with the multi class log-loss shown in Eq. (1)[1].

Fourteen (14) multi class document classification problems which their properties are summarized in Table 1 were used. The 20 Newsgroups dataset contains approximately $20,000$ newsgroup messages. The two Amazon datasets contain product reviews, and given a review, the task is to predict one of seven product categories, or a subset of three categories. The tasks of the seven Enron datasets is to automatically sort emails into one of ten folders. The three tasks based on the New York Times corpus are to predict the desk that produced the story (Financial, Sports, etc.), the online section to which the article was posted, and the section in which the article was printed. Finally, the Reuters corpus (RCV1-v2) contains a subset of $5,000$ newswire stories that should be labeled with one of four general topics: corporate, economics, government and markets. Additional details can be found in a recent paper [Crammer *et al.*, 2009]. The two right columns show the mean and standard deviation of numbers of examples per class. Clearly some datasets are well balanced while other are far from being balanced.

We first compared the algorithms (or regularizations) in terms of classification performance. The value of the trade-off parameter λ was set using a random split of the data into a training set consists of 80% of the examples and a test set with the remaining 20%. This partition was used for the optimal λ selection. The goal of the process was to find the parameters with the optimal performance. The final results are based on 8-fold cross validation.

Performance is evaluated using Macro-F1 which averages F1 over harmonic mean of precision and recall per class. We also use macro-precision and macro-recall, each are averages of precision and recall over classes. Additionally, we evaluate the sparsity of a model by the fraction of zero-elements in it. Higher F1 indicates better prediction performance, while higher sparsity indicates smaller models.

[1]We have also tried the multi class hinge loss mentioned above, yet performance in general was inferior to the log-loss and thus it was omitted.

Dataset	Ex. #	Features	Cl. #	Ex. per class	
				Mean	STD
20 News	18828	252122	20	941	97
Amazon 7	13580	686724	7	1940	0
Amazon 3	38781	1876019	3	12927	0
Enron bec	751	7134	10	75	37
Enron far	3020	13561	10	302	290
Enron kam	3172	18441	10	317	194
Enron kit	2345	15688	10	235	163
Enron lok	1966	16012	10	197	278
Enron san	863	10921	10	86	108
Enron wil	2542	8816	10	254	487
NYT desk	10000	114534	26	385	703
NYT online	10000	114534	25	400	699
NYT section	10000	114534	20	500	827
Reuters 4	5000	268170	4	1250	725

Table 1: Multi class datasets

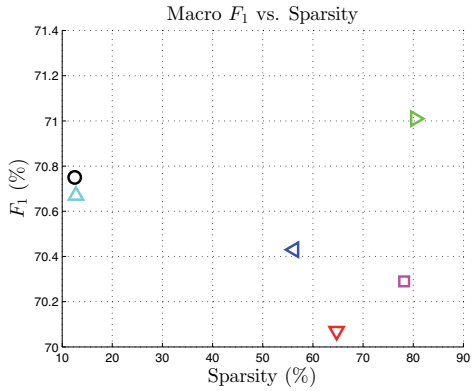
Fig. 1(a) shows the Macro-F1 vs sparsity for the six regularizations, averaged over all datasets. Higher points indicate better prediction performance, while points to the right, indicate sparser models. Both $\ell_{2,2}$ and $\ell_{2,1;2}$, yield similar performance and, as expected, very low sparsity. The performance of $\ell_{1,1}$, $\ell_{1,2;1}$ and $\ell_{2,1;1}$ is lower, yet yield sparser models. The best algorithm both in terms performance and sparsity is $\ell_{1,2;2}$. This is surprising as we evaluate sparsity using the entire sparsity which $\ell_{1,1}$ is minimizing.

To better understand the performance of each algorithm we plot in Fig. 1(b) the precision vs recall. Regularization with $\ell_{2,2}$ norm yields relatively high precision and recall. The common sparsity regularization $\ell_{1,1}$ is worse both in terms of recall and precision. Our proposed sparsity regularization $\ell_{1,2;2}$ has lower precision than $\ell_{2,2}$ (yet better than $\ell_{1,1}$), but achieves the highest recall, which allows it to have the highest F1 altogether.

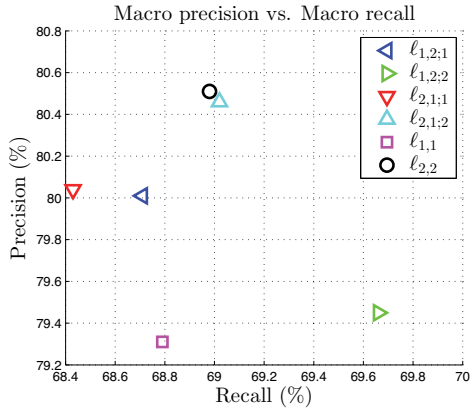
We next analyze the structure of the models over different rates of sparsity, as it may be the case that we are limited by sparsity constraints. We trained 128 models for all five regularizations (excluding $\ell_{2,2}$) using lambda values in the range of $10^{-5} - 10^2$.

For each of the models we computed class and feature sparsity statistics, specifically the standard deviation of each group sparsity. First, we compute the class sparsity by computing the non zero rates of class terms per feature t: $\frac{1}{c}\sum_{s=1}^{c} I\left[\boldsymbol{\omega}_{t,c} \neq 0\right]$ and then taking the standard deviation of these quantities, denoted by STD_c. Second, similarly, we compute the feature sparsity by computing the non zero rates of feature terms per class s: $\frac{1}{d}\sum_{t=1}^{d} I\left[\boldsymbol{\omega}_{t,c} \neq 0\right]$, and then taking the standard deviation of these quantities, denoted by STD_f. A low value of STD_c indicates that for all features the amount of class sparsity is close to uniform. That is, the number of non-zero elements per row is close to a constant. Similarly, a low STD_f indicates a consistent feature sparsity.

Results for three datasets are given in Fig 2. The left panels show STD_f vs total amount of non zero rates for three datasets, and the right panels show the STD_c vs total amount of non zero rates for the same datasets. In all plots the STD goes to zero when the non zero rates go to zero (high spar-

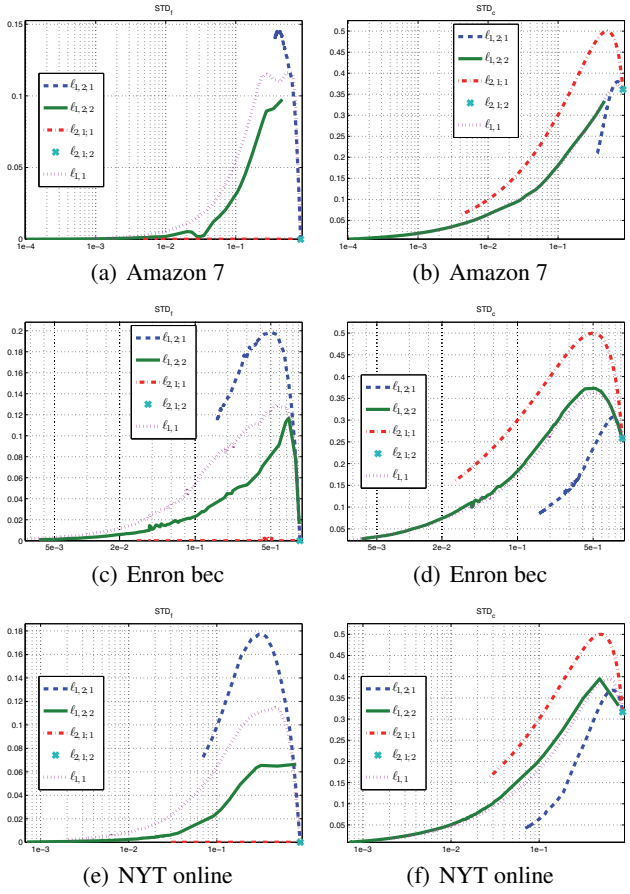


(a) Macro F_1 vs. Sparsity(%)



(b) Macro Precision vs. Macro Recall

Figure 1: Summary of performance results on 14 multi class tasks



(a) Amazon 7 (b) Amazon 7

(c) Enron bec (d) Enron bec

(e) NYT online (f) NYT online

Figure 2: Group sparsity STD vs fraction of non-zero elements (log scale). Left: Feature sparsity STD. Right: Class sparsity STD.

sity), as more and more entries of the model are set to zero.

Focusing on the left panels we see that STD_f is zero for $\ell_{2,1;1}$, as this regularization either does not enforce a sparsity of a feature, or cancels the entire feature altogether, and thus there is a constant feature sparsity for all classes. The second best (or uniform) regularization is $\ell_{1,2;2}$, then $\ell_{1,1}$, and $\ell_{1,2;1}$ has the least consistent feature sparsity, as it enforces large amount of zero entries per feature, yet the number of classes is no more than 50 (the options for terms removal are limited).

Moving to the right panels we observe that $\ell_{1,2;1}$ has the lowest STD_c, then, $\ell_{1,1}$ and $\ell_{1,2;2}$ are close to each other, and $\ell_{2,1;1}$ is with the highest STD_c. In both measures, $\ell_{1,1}$ STD values are mediocre compared to other regularizations' values which fits its lack of specific sparsity structure. As expected, $\ell_{1,2;1}$ has a consistent class sparsity (low STD_c) and $\ell_{1,2;2}$ has a consistent feature sparsity (low STD_f). However, $\ell_{1,2;1}$ has a low feature consistency (high STD_f) while $\ell_{1,2;2}$ has a medium class consistency (very close to $\ell_{1,1}$). This can be explained, as before, by the fact that $c \ll d$, which enables scattered non zero terms for each of the classes, allowing a higher chance for features with a consistent class sparsity. $\ell_{2,1;2}$ yields only dense results, hence not relevant for this analysis.

Similar behavior was evident in all other datasets.

6 Conclusions and Future Work

This work presents a novel approach for multi class problems, proposing an individual feature selection, which is formulated by the $\ell_{1,2;2}$ norm regularization. This approach was thoroughly investigated and compared with the common $\ell_{2,2}$, $\ell_{1,1}$ and $\ell_{2,1}$ regularizations, examining performance and sparsity pattern results. The empirical study, conducted on 14 multi class datasets, demonstrates the superiority of the $\ell_{1,2;2}$ regularization for the multi class problems, outperforming other regularizations. Interestingly, results for $\ell_{1,2;2}$ are not only better, but also sparser than all other regularizations, including the general sparsity promoting $\ell_{1,1}$. It was also shown that sparsity patterns fitted the expectations with consistent feature sparsity results for $\ell_{1,2;2}$. In addition, theoretical guarantees were proved using robustness and Rademacher analysis. These guarantees supply theoretical reasoning for choosing $\ell_{1,2;2}$ norm, given a specific prior knowledge.

Future work may analyze and find criteria for performance advantage of the proposed regularization and further investigate specific loss functions and types of datasets, as well as additional types of problems, e.g. multi task. Another possible direction is exploring combinations (sums) of different regularizations to obtain a mixed effect.

References

[Bach *et al.*, 2011] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Convex optimization with sparsity-inducing norms. In *Optimization for Machine Learning*. MIT Press, 2011.

[Bakin, 1999] Sergey Bakin. Adaptive regression and model selection in data mining problems, 1999. Thesis (Ph.D.)–Australian National University, 1999.

[Bartlett and Mendelson, 2003] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, March 2003.

[Benedek and Panzone, 1961] A Benedek and Panzone. R.: The space lp, with mixed norm. *Duke Math. J*, (28):301–324, 1961.

[Boser *et al.*, 1992] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, 1992.

[Bradley and Bagnell, 2009] David Bradley and J. Andrew (Drew) Bagnell. Convex coding. Technical Report CMU-RI-TR-09-22, Robotics Institute, Pittsburgh, PA, May 2009.

[Combettes and Pesquet, 2011] Patrick Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In H.H. Bauschke, R.S. Burachik, P.L. Combettes, V. Elser, D.R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, volume 20, pages 273–297, 1995.

[Crammer and Singer, 2001] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, December 2001.

[Crammer *et al.*, 2009] Koby Crammer, Mark Dredze, and Alex Kulesza. Multi-class confidence weighted algorithms. In *EMNLP*, pages 496–504, 2009.

[Duchi and Singer, 2009a] John Duchi and Yoram Singer. Boosting with structural sparsity. In *ICML*, ICML '09, 2009.

[Duchi and Singer, 2009b] John Duchi and Yoram Singer. Efficient learning using forward-backward splitting. In *NIPS*, pages 495–503, 2009.

[Fornasier and Rauhut, 2008] Massimo Fornasier and Holger Rauhut. Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, 46(2):577–613, 2008.

[Hoerl and Kennard, 1970] A E Hoerl and R W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, (12):55–67, 1970.

[Hoerl, 1962] A E Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58(3):54–59, 1962.

[Kowalski and Torrésani, 2008] Matthieu Kowalski and Bruno Torrésani. Sparsity and persistence: mixed norms provide simple signal models with dependent coefficients. *Signal Image and Video Processing*, 3(3):251–264, 2008.

[Kowalski *et al.*, 2009] Matthieu Kowalski, Marie Szafranski, and Liva Ralaivola. Multiple indefinite kernel learning with mixed norm regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 545–552, 2009.

[Kowalski, 2009] Matthieu Kowalski. Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303 – 324, 2009.

[Lee *et al.*, 2004] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81, 2004.

[Mairal *et al.*, 2010] Julien Mairal, Rodolphe Jenatton, Guillaume Obozinski, and Francis Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems 23*, pages 1558–1566. 2010.

[Massart, 2000] Pascal Massart. Some Applications of Concentration Inequalities to Statistics. *Annales de la Faculté des Sciences de Toulouse*, IX(2):245–303, 2000.

[Moreau, 1962] J Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, (255), 1962.

[Mukherjee *et al.*, 2002] Sayan Mukherjee, Ryan Rifkin, and Tomaso Poggio. Regression and classification with regularization, 2002.

[Szafranski *et al.*, 2008] Marie Szafranski, Yves Grandvalet, and Pierre Morizet-Mahoudeaux. Hierarchical penalization. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, 2008.

[Teschke and Ramlau, 2007] Gerd Teschke and Ronny Ramlau. An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector-valued regimes and an application to color image inpainting. *Inverse Problems*, 23(5):1851, 2007.

[Tibshirani, 1996] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological*, (58), 1996.

[Tikhonov and Arsenin, 1977] A N Tikhonov and V Y Arsenin. Solutions of ill posed problems, 1977.

[Weston and Watkins, 1998] J Weston and C Watkins. Multi-class support vector machines. *Pattern Recognition*, (CSD-TR-98-04):0–9, 1998.

[Xu *et al.*, 2009] H Xu, C Caramanis, and S Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, (10):1485–1510, 2009.

[Yuan and Lin, 2006] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[Zhao *et al.*, 2009] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *ANNALS OF STATISTICS*, 37:3468, 2009.