# EPIGRAPHICAL PROXIMAL PROJECTION FOR SPARSE MULTICLASS SVM

*Giovanni Chierchia*[1]*, Nelly Pustelnik*[2]*, Jean-Christophe Pesquet*[3]*, and Béatrice Pesquet-Popescu*[1]

[1] Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, 75014 Paris, France
[2] Laboratoire de Physique - ENSL, UMR CNRS 5672, F-69007 Lyon, France
[3] Université Paris-Est, LIGM, UMR CNRS 8049, 77454 Marne-la-Vallée, France
Contact: chierchi@telecom-paristech.fr

## ABSTRACT

Sparsity inducing penalizations are useful tools in variational methods for machine learning. In this paper, we design a learning algorithm for multiclass support vector machines that allows us to enforce sparsity through various nonsmooth regularizations, such as the mixed $\ell_{1,p}$-norm with $p \geq 1$. The proposed constrained convex optimization approach involves an epigraphical constraint for which we derive the closed-form expression of the associated projection. This sparse multiclass SVM problem can be efficiently implemented thanks to the flexibility offered by recent primal-dual proximal algorithms. Experiments carried out for handwritten digits demonstrate the interest of considering nonsmooth sparsity-inducing regularizations and the efficiency of the proposed epigraphical projection method.

***Index Terms***— Convex optimization, SVM, sparsity, epigraphical projection, proximal methods

## 1. INTRODUCTION

Support vector machines (SVMs) have gained much popularity in solving large-scale classification problems, thanks to their excellent performance and their ability to efficiently deal with large datasets. In order to predict the class $z \in \{1, \ldots, K\}$ that best matches an observation $u \in \mathbb{R}^N$ (e.g. a signal, an image or a graph), SVMs rely on a discriminant function $D\colon \mathbb{R}^N \times \{1, \ldots, K\} \mapsto \mathbb{R}$ which is built from a set of $L$ input-output pairs $\mathcal{S} = \big\{(\mathrm{u}^{(\ell)}, z^{(\ell)}) \in \mathbb{R}^N \times \{1, \ldots, K\} \mid \ell \in \{1, \ldots, L\}\big\}$. This function aims at partitioning the observation space into $K$ regions (one for each expected class) and it is estimated so that the separating hyperplanes maximize the distance to the nearest training point of any class. Such a maximum-margin principle was originally formulated in [1] for $K = 2$ and then generalized in [2] for $K \geq 2$.

Most of the applications considered in the literature deal with a large amount of training data [3, 4] or a huge (even infinite) number of classes [5]. Consequently, the major difficulty encountered in this kind of applications stems from the computational cost. On the other hand, in some applications, only a small number of training data is available. This is undoubtedly true in medical contexts, where the goal is to classify a patient as "being in good health", "being contaminated", or "being infected", but the verified cases of infected patients might be just a few. In such applications, the lack of training data may lead to the so-called *overfitting*, eventually leading to a prediction which is too strongly tailored to the particularities of the training set and poorly generalizes to new data.

Formally, the discriminant function is assumed to be linear in some combined feature representation of inputs and outputs [6]. This assumption leads to

$$\big(\forall (\mathrm{u}, z) \in \mathbb{R}^N \times \{1, \ldots, K\}\big) \qquad D(\mathrm{u}, z) = \mathrm{x}^\top \Psi(\mathrm{u}, z),$$

where, for every $z \in \{1, \ldots, K\}$, $\Psi(\cdot, z)\colon \mathbb{R}^N \to \mathbb{R}^{MK}$ is such that $\mathrm{x}^\top \Psi(\mathrm{u}, z) = (\mathrm{x}^{(z)})^\top \phi(\mathrm{u})$, the function $\phi\colon \mathbb{R}^N \to \mathbb{R}^M$ denotes a mapping[1] from the input space $\mathbb{R}^N$ onto an arbitrary feature space $\mathbb{R}^M$, and $\mathrm{x} = (\mathrm{x}^{(z)})_{1 \leq z \leq K} \in \mathbb{R}^{MK}$ denotes the vector to be estimated, block decomposed into vectors $\mathrm{x}^{(z)} \in \mathbb{R}^M$ with $z \in \{1, \ldots, K\}$.

**Related works**. The multiclass SVM proposed in [2] amounts to solving the following convex optimization problem

$$\underset{(\mathrm{x}, \xi) \in \mathbb{R}^{MK} \times \mathbb{R}^L}{\text{minimize}} \quad \frac{1}{2}\|\mathrm{x}\|^2 + \lambda \sum_{\ell=1}^{L} \xi^{(\ell)} \quad \text{subj. to}$$
$$\begin{cases} (\forall \ell \in \{1, \ldots, L\})(\forall j \in \{1, \ldots, K-1\}) \\ \qquad\qquad \mathrm{x}^\top w^{(\ell, j)} \geq 1 - \xi^{(\ell)}, \\ (\forall \ell \in \{1, \ldots, L\}) \qquad \xi^{(\ell)} \geq 0, \end{cases} \quad (1)$$

where $\xi = (\xi^{(\ell)})_{1 \leq \ell \leq L}$ is the vector of slack variables, $\lambda > 0$ is a regularization constant, while for every $\ell \in \{1, \ldots, L\}$,

$$(w^{(\ell, j)})_{1 \leq j \leq K-1} = \Big(\Psi(\mathrm{u}^{(\ell)}, z) - \Psi(\mathrm{u}^{(\ell)}, z^{(\ell)})\Big)_{z \in \{1, \ldots, K\} \setminus \{z^{(\ell)}\}}.$$

The above problem was solved in [2] by using standard Lagrangian duality techniques. While the dual formulation brings in several advantages (e.g. the kernel trick [7]), the problem

[1] The mapping $\phi$ allows one to fit the maximum-margin hyperplanes in a transformed feature space, where the observations are more likely to be linearly separable.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

size becomes prohibitive when the number of classes is high. Hence, recent works [5, 8] proposed to approximate the dual problem using cutting plane approaches, in order to address scenarios with thousands or even an infinite number of classes.

Since the features are not equally informative, a common solution to prevent overfitting consists of imposing a sparsity condition on the vector $\mathrm{x}$. In this respect, the $\boldsymbol{\ell}_1$-norm and, more generally, the mixed $\boldsymbol{\ell}_{1,p}$-norm have proven to be useful in several machine-learning applications [9, 10, 11]. However, when a nonsmooth penalty is substituted for the $\boldsymbol{\ell}_2$-norm, the dual formulation becomes non trivial. For this reason, when sparse classification is proposed, the maximum-margin principle is equivalently formulated by using the *hinge loss function* [12, 13, 14, 15]. A different approach for sparse classification consists of replacing the hinge loss with other types of loss functions, such as the logistic loss [16, 13, 17]. All these solutions aim at simplifying the optimization procedure, but they do not solve rigorously (1) with a sparse penalization.

**Contributions**. In this work, we propose an efficient solution to exactly solve (1) in the case when the $\boldsymbol{\ell}_2$-norm is replaced by any convex, lower semi-continuous, and proper function $g$ from $\mathbb{R}^{MK}$ to $]-\infty,+\infty]$. The only assumption required by our method is that the proximity operator [18] of $g$ can be calculated explicitly. This is certainly the case for the mixed $\boldsymbol{\ell}_{1,p}$-norm with $p \in \{1, 2, +\infty\}$ [19, 20, 21]. The paper is organized as follows. In Section 2, we formulate the multiclass problem in terms of nonlinear epigraphical constraints, in Section 3 we provide the proximal tools and the epigraphical projection needed to solve the proposed problem, and in Section 4 we compare our solution with the conventional $\boldsymbol{\ell}_2$-SVM on a standard database.

**Notation**. $\Gamma_0(\mathbb{R}^N)$ denotes the set of proper, lower semicontinuous, convex functions from $\mathbb{R}^N$ to $]-\infty,+\infty]$. The epigraph of $\varphi \in \Gamma_0(\mathbb{R}^N)$ is the nonempty closed convex subset of $\mathbb{R}^N \times \mathbb{R}$ defined as $\operatorname{epi}\varphi = \big\{(y,\zeta) \in \mathbb{R}^N \times \mathbb{R} \;\big|\; \varphi(y) \leq \zeta\big\}$. For every $y \in \mathbb{R}^N$, the proximity operator of $\varphi$ is $\operatorname{prox}_{\varphi}(y) = \operatorname{argmin}_{u\in\mathbb{R}^N} \|u-y\|^2 + \varphi(u)$ and the projection onto a nonempty closed convex subset $C \subset \mathbb{R}^N$ is $P_C(y) = \operatorname{prox}_{\iota_C}(y) = \operatorname{argmin}_{u\in C} \|u-y\|^2$, where $\iota_C$ is the indicator function of $C$, equal to 0 on $C$ and $+\infty$ otherwise.

## 2. SPARSE MULTICLASS SVM

We extend Problem (1) by replacing the $\boldsymbol{\ell}_2$-norm regularization with a generic function $g \in \Gamma_0(\mathbb{R}^{MK})$ and by considering a constrained structural-risk minimization. To do so, for every $\ell \in \{1,\dots,L\}$, we introduce the function

$$(\forall \mathrm{y}^{(\ell)} = (y^{(\ell,j)})_{1\leq j\leq K-1} \in \mathbb{R}^{(K-1)}) \qquad h^{(\ell)}(\mathrm{y}^{(\ell)}) = \max_{1\leq j\leq K-1} y^{(\ell,j)} + \mu, \quad (2)$$

where $y^{(\ell,j)} = -\mathrm{x}^\top w^{(\ell,j)}$ and $\mu$ is a positive constant that allows us to model the margin-rescaling criterion in [22]. Consequently, in order to estimate the vector $\mathrm{x}$ from the training data in $S$, we aim at solving the convex problem:

$$\underset{(\mathrm{x},\xi)\in\mathbb{R}^{MK}\times\mathbb{R}^L}{\text{minimize}} \quad g(\mathrm{x}) \quad \text{subj. to} \quad \begin{cases} \xi^{(1)} + \cdots + \xi^{(L)} \leq \eta, \\ (\forall \ell \in \{1,\dots,L\}) \qquad h^{(\ell)}(\mathrm{y}^{(\ell)}) \leq \xi^{(\ell)}, \\ (\forall \ell \in \{1,\dots,L\}) \qquad \xi^{(\ell)} \geq 0, \end{cases} \quad (3)$$

where $\eta$ is a positive constant. Note that the above decomposition yields the same reformulation of Problem (1) as considered in [2], except for the function $g$ and the half-space constraint over the slack vector. Indeed, the above constrained formulation is equivalent to Problem (1) for some specific values of $\eta$ and $\lambda$, but the constrained one allows us to control more easily the effect of slack variables. The advantage of the constrained formulation is that the choice of $\eta$ may be easier, since it is directly related to the properties of training data.

The function $g$ is chosen so as to prefer a *simple* solution rather than a *complex* one. This condition is typically achieved by promoting a sparse solution. Sparsity can be enforced with different regularization functions. A popular example is the $\boldsymbol{\ell}_1$-norm [9], which is known to induce sparsity: the solution will have a number of coefficients exactly equal to zero, depending on the strength of the regularization. Another example is given by the mixed $\boldsymbol{\ell}_{1,p}$-norm [9], defined for each $\mathrm{x} = (\mathrm{x}^{(z)})_{1\leq z\leq K} \in \mathbb{R}^{MK}$ which is block-decomposed, for every $z \in \{1,\dots,K\}$, as $\mathrm{x}^{(z)} = [\mathrm{x}^{(z,1)^\top} \dots \mathrm{x}^{(z,B)^\top}]^\top \in \mathbb{R}^M$:

$$\|\mathrm{x}\|_{1,p} = \sum_{z=1}^{K}\sum_{b=1}^{B} \|\mathrm{x}^{(z,b)}\|_p. \quad (4)$$

The mixed-norm is known to induce *block-sparsity*: the solution is partitioned into groups and the variables of each group are ideally either all zeros or all non-zeros. In this context, the exponent values $p = 2$ or $p = +\infty$ are the most popular choices. In particular, the $\boldsymbol{\ell}_{1,\infty}$-norm tends to favour solutions with many components of equal magnitude.

## 3. ALGORITHMIC SOLUTION

Within the proposed constrained optimization framework, a possible reformulation of Problem (3) is the following:

$$\underset{(\mathrm{x},\xi)\in\mathbb{R}^{MK}\times V}{\text{minimize}} \quad g(\mathrm{x}) \quad \text{subj. to} \quad (W\mathrm{x},\xi) \in E, \quad (5)$$

where $W \in \mathbb{R}^{L(K-1)\times MK}$ is the linear operator such that

$$W\mathrm{x} = \mathrm{y} = (\mathrm{y}^{(\ell)})_{1\leq\ell\leq L}, \quad (6)$$

the set $V$ denotes the simplex

$$V = \big\{\xi \in [0,+\infty[^L \;\big|\; \xi^{(1)} + \cdots + \xi^{(L)} \leq \eta\big\}, \quad (7)$$

and the set $E$ is a collection of epigraphs

$$E = \big\{(\mathrm{y},\xi) \in \mathbb{R}^{L(K-1)} \times \mathbb{R}^L \;\big|\; (\forall \ell \in \{1,\dots,L\}) \quad (\mathrm{y}^{(\ell)},\xi^{(\ell)}) \in \operatorname{epi} h^{(\ell)}\big\}. \quad (8)$$

### 3.1. Epigraphical projection

The advantage of the epigraphical decomposition is that the projections $P_E$ and $P_V$ onto $E$ and $V$ have closed-form expressions. Indeed, the projection $P_V$ is given in [23], while the projection $P_E$ is block-decomposed as

$$P_E(\mathrm{y},\xi) = \Big(P_{\operatorname{epi} h^{(\ell)}}(\mathrm{y}^{(\ell)},\xi^{(\ell)})\Big)_{1\leq\ell\leq L}, \tag{9}$$

where, for every $(\mathrm{y}^{(\ell)},\xi^{(\ell)}) \in \mathbb{R}^{K-1}\times\mathbb{R}$, $P_{\operatorname{epi} h^{(\ell)}}(\mathrm{y}^{(\ell)},\xi^{(\ell)})$ denotes the solution of

$$\min_{(\mathsf{p}^{(\ell)},\theta^{(\ell)})\in\operatorname{epi} h^{(\ell)}} \|\mathsf{p}^{(\ell)}-\mathrm{y}^{(\ell)}\|^2 + (\theta^{(\ell)}-\xi^{(\ell)})^2 \tag{10}$$

which is equivalent to find

$$\min_{\theta^{(\ell)}\in\mathbb{R}} \Big\{(\theta^{(\ell)}-\xi^{(\ell)})^2 + \min_{\substack{p^{(\ell,1)}\leq\theta^{(\ell)}-\mu\\ \vdots \\ p^{(\ell,K-1)}\leq\theta^{(\ell)}-\mu}} \|\mathsf{p}^{(\ell)}-\mathrm{y}^{(\ell)}\|^2\Big\}.$$

For every $\theta^{(\ell)}\in\mathbb{R}$, the inner minimization is achieved when, for every $j\in\{1,\dots,K-1\}$, $p^{(\ell,j)}$ is the projection of $y^{(\ell,j)}$ onto the real interval $]-\infty,\theta^{(\ell)}-\mu]$. Consequently, Problem (10) reduces to

$$\min_{\theta^{(\ell)}\in\mathbb{R}} \Big\{(\theta^{(\ell)}-\xi^{(\ell)})^2 + \sum_{j=1}^{K-1}(\max\{y^{(\ell,j)}+\mu-\theta^{(\ell)},0\})^2\Big\}$$

which is also equivalent to calculate, at the point $\xi^{(\ell)}$, the proximity operator [18] of the following convex function:

$$(\forall v\in\mathbb{R}) \qquad \varphi(v) = \frac{1}{2}\sum_{j=1}^{K-1}(\max\{y^{(\ell,j)}+\mu-v,0\})^2. \tag{11}$$

The closed form expression of this proximity operator is given by [24, Proposition II.8] and it is summarized in the following proposition.

**Proposition 3.1.** *Let $\big(\nu^{(\ell,j)}\big)_{1\leq j\leq K-1}$ be a sequence obtained by sorting $\big(y^{(\ell,j)}+\mu\big)_{1\leq j\leq K-1}$ in ascending order, and set $\nu^{(\ell,0)}=-\infty$ and $\nu^{(\ell,K)}=+\infty$. Then, for every $(\mathrm{y}^{(\ell)},\xi^{(\ell)})\in\mathbb{R}^{K-1}\times\mathbb{R}$, the projection $P_{\operatorname{epi} h^{(\ell)}}(\mathrm{y}^{(\ell)},\xi^{(\ell)}) = (\mathsf{p}^{(\ell)},\theta^{(\ell)})$ is such that $\mathsf{p}^{(\ell)}=(p^{(\ell,j)})_{1\leq j\leq K-1}$ with, for every $j\in\{1,\dots,K-1\}$,*

$$p^{(\ell,j)} = \begin{cases} y^{(\ell,j)}, & \text{if } y^{(\ell,j)}\leq\theta^{(\ell)}-\mu,\\ \theta^{(\ell)}-\mu, & \text{otherwise}, \end{cases} \tag{12}$$

*and*

$$\theta^{(\ell)} = \frac{1}{K-\overline{j}^{(\ell)}+1}\left(\xi^{(\ell)}+\sum_{j=\overline{j}^{(\ell)}}^{K-1}\nu^{(\ell,j)}\right), \tag{13}$$

*where $\overline{j}^{(\ell)}$ is the unique integer in $\{1,\dots,K\}$ such that*

$$\nu^{(\ell,\overline{j}^{(\ell)}-1)} < \theta^{(\ell)} \leq \nu^{(\ell,\overline{j}^{(\ell)})} \tag{14}$$

*(with the convention $\sum_{j=K}^{K-1}\cdot = 0$).*

### 3.2. Proposed algorithm

The solution of (5) requires an efficient algorithm for dealing with nonsmooth functions. We resort here to proximal algorithms. Among the large panel of existing proximal algorithms [25, 26, 27], we consider the primal-dual M+LFBF algorithm recently proposed in [27], which is able to address general convex optimization problems involving nonsmooth functions and linear operators without requiring any matrix inversion. This algorithm is able to solve:

$$\underset{\mathrm{v}\in\mathcal{H}}{\operatorname{minimize}} \;\; \phi(\mathrm{v}) + \sum_{i=1}^{Q}\psi_i(T_i\mathrm{v}) \tag{15}$$

where $\mathcal{H}$ is a real Hilbert space, $\phi\colon \mathcal{H}\mapsto\,]-\infty,+\infty]$ is a proper convex lower-semicontinuous function, for every $i\in\{1,\dots,Q\}$, $T_i\colon\mathcal{H}\mapsto\mathbb{R}^{S_i}$ is a bounded linear operator and $\psi_i\colon\mathbb{R}^{S_i}\mapsto\,]-\infty,+\infty]$ is a proper convex lower-semicontinuous function.

Our minimization problem fits nicely into this framework by setting $\mathcal{H}=\mathbb{R}^{MK}\times\mathbb{R}^L$, $\mathrm{v}=(\mathrm{x},\xi)$, $Q=1$ and $S_1 = L(K-1)+MK$. The linear operator is

$$T_1 = \begin{bmatrix} W & 0\\ 0 & \mathrm{Id}\end{bmatrix}$$

and the functions are the following ones:

$$\begin{aligned} &(\forall(\mathrm{x},\xi)\in\mathbb{R}^{MK}\times\mathbb{R}^L) && \phi(\mathrm{x},\xi) &&= g(\mathrm{x})+\iota_V(\xi),\\ &(\forall(\mathrm{y},\xi)\in\mathbb{R}^{L(K-1)}\times\mathbb{R}^L) && \psi_1(\mathrm{y},\xi) &&= \iota_E(\mathrm{y},\xi).\end{aligned}$$

The iterations associated with Problem (3) are summarized in Algorithm 1, where the sequence $(\mathrm{x}^{[i]})_{i\in\mathbb{N}}$ is guaranteed to converge to a solution to (3), provided that such a solution exists [27].

**Algorithm 1** M+LFBF for solving Problem (3)

Initialization
- $(\mathrm{y}^{[0]},\nu^{[0]})\in\mathbb{R}^{L(K-1)}\times\mathbb{R}^L$
- $(\mathrm{x}^{[0]},\xi^{[0]})\in\mathbb{R}^{MK}\times\mathbb{R}^L$
- $\beta=\max\{\|W\|,1\}$
- $\epsilon\in]0,1/(\beta+1)[$

For $i=0,1,\dots$
- $\gamma_i\in[\epsilon,(1-\epsilon)/\beta]$
- $\Big(\mathrm{p}^{[i]},\rho^{[i]}\Big) = \Big(\operatorname{prox}_{\gamma_i g}(\mathrm{x}^{[i]}-\gamma_i W^\top\mathrm{y}^{[i]}), P_V(\xi^{[i]}-\gamma_i\nu^{[i]})\Big)$
- $\Big(\widehat{\mathrm{y}}^{[i]},\widehat{\nu}^{[i]}\Big) = \Big(\mathrm{y}^{[i]},\nu^{[i]}\Big)+\gamma_i\Big(W\,\mathrm{x}^{[i]},\xi^{[i]}\Big)$
- $\Big(\mathrm{a}^{[i]},\alpha^{[i]}\Big) = \Big(\widehat{\mathrm{y}}^{[i]},\widehat{\nu}^{[i]}\Big)-\gamma_i P_E\Big(\widehat{\mathrm{y}}^{[i]}/\gamma_i,\widehat{\nu}^{[i]}/\gamma_i\Big)$
- $\Big(\mathrm{y}^{[i+1]},\nu^{[i+1]}\Big) = \Big(\mathrm{a}^{[i]},\alpha^{[i]}\Big)+\gamma_i\Big(W(\mathrm{p}^{[i]}-\mathrm{x}^{[i]}),\rho^{[i]}-\xi^{[i]}\Big)$
- $\Big(\mathrm{x}^{[i+1]},\xi^{[i+1]}\Big) = \Big(\mathrm{p}^{[i]},\rho^{[i]}\Big)-\gamma_i\Big(W^\top(\mathrm{a}^{[i]}-\mathrm{y}^{[i]}),\alpha^{[i]}-\nu^{[i]}\Big)$

### 3.3. Formulation based on linear constraints

At this point, we would like to emphasize that more standard formulations of Problem (3) are possible. For example, one may naturally think of introducing a vector $\zeta \in \mathbb{R}^{L(K-1)}$ and rewrite the inequalities in Problem (3) in terms of an extended number of linear constraints

$$\begin{cases} (\forall \ell \in \{1,...,L\})(\forall j \in \{1,...,K-1\})\ \zeta^{(\ell,j)} \geq 0, \\ (\forall \ell \in \{1,...,L\})(\forall j \in \{1,...,K-1\})\ y^{(\ell,j)} + \mu \leq \zeta^{(\ell,j)}, \\ (\forall \ell \in \{1,...,L\}) \quad \zeta^{(\ell,1)} = \cdots = \zeta^{(\ell,K-1)}, \\ \displaystyle\sum_{\ell=1}^{L} \sum_{j=1}^{K-1} \zeta^{(\ell,j)} \leq (K-1)\,\eta. \end{cases} \tag{16}$$

In this regard, we will show in Section 4 that the proposed epigraphical reformulation converges much faster than the solution based on linear constraints.

## 4. NUMERICAL RESULTS

We perform our experimental analysis with an example of handwritten digit classification. More precisely, we consider the MNIST database,[2] which contains a large number of grayscale images displaying handwritten digits from 0 to 9. The images were size-normalized to fit into a $20 \times 20$ pixel box, and then centered in a $28 \times 28$ image [28]. The database is organized in 60000 training images and 10000 test images.

In our experiments, we scaled the image dynamics range to the interval $[0, 1]$ by dividing the pixel intensities by 255. Moreover, we selected $L$ image-class pairs $(\mathrm{u}^{(\ell)}, \mathrm{z}^{(\ell)})_{1 \leq \ell \leq L} \in \mathbb{R}^{28^2} \times \mathcal{Z}$ from the training set, with $\mathcal{Z} = \{1, \ldots, 10\}$, and we defined the mapping $\phi : \mathbb{R}^{28^2} \mapsto \mathbb{R}^M$ by resorting to the scattering convolution network recently proposed in [29], using $\overline{m} = 2$ wavelet layers scaled up to $2^J = 4$, yielding $M = 15876$.

We evaluated the impact of the regularization over the performance obtained with the considered multiclass SVM and we compared it with the sparse multinomial logistic regression [17]. For SVM, we set $\mu \equiv 1$ in (2) and, for the regularization, we considered the $\ell_2$-norm and the $\ell_{1,\infty}$-norm, which recently gained much attention in learning tasks [10, 11]. To evaluate the quality of the estimated vector $\mathrm{x} \in \mathbb{R}^{10M}$, we collected in Table 1 the misclassification errors obtained by evaluating the prediction $d_{\mathrm{x}}(\mathrm{u}) = \mathrm{argmax}_{z \in \{1,\ldots,K\}}\, \mathrm{x}^\top \Psi(\mathrm{u}, z)$ on the 10000 test images. The results indicate that the block-sparse $\ell_{1,\infty}$-norm regularization makes a significant difference in the case when a few examples are available for training.

In Fig. 1, we show that the epigraphical approach (solid blue line) leads to a faster convergence (about 4 times) than a more standard technique for handling linear constraints (dashed red line). The results refer to the case $L = 100$ with

[2] available at `http://yann.lecun.com/exdb/mnist`

| $L/K$ | $\ell_2$-SVM | $\ell_{1,\infty}$-SVM | $\ell_1$-logit [17] |
|---|---|---|---|
| 3 | 27.06 % | **25.64** % | 28.14 % |
| 5 | 16.32 % | **13.59** % | 15.48 % |
| 10 | 11.00 % | **9.40** % | 10.42 % |
| 15 | 10.12 % | **7.68** % | 8.75 % |
| 20 | 7.78 % | **5.67** % | 6.18 % |
| 30 | 6.48 % | **5.46** % | 5.69 % |
| 50 | 4.22 % | **3.73** % | 3.92 % |
| 100 | 3.69 % | **3.13** % | 3.34 % |

**Table 1**. Classification errors obtained by using different regularizations within the considered multiclass SVM.

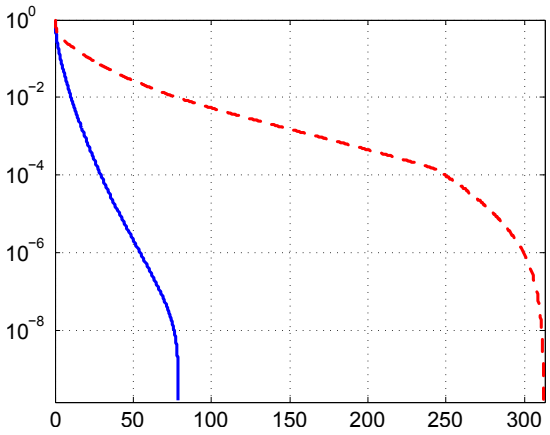


**Fig. 1**. Relative error $\|x^{[n]} - x^{[\infty]}\|/\|x^{[\infty]}\|$ vs computational time (in seconds), where $x^{[\infty]}$ denotes the solution computed after a large number (10000) of iterations. Red line: approach with linear constraints. Blue line: epigraphical approach.

$\ell_2$-norm regularization. Our codes were completely developed in MATLAB and all the programs executed on an Intel Xeon CPU X5690 at 3.47 GHz and 24 GB of RAM.

## 5. CONCLUSIONS

We have proposed a new epigraphical technique for solving constrained convex optimization problems arising in machine learning with support vector machines. In particular, the epigraphical splitting allows us to handle the multiclass maximum-margin loss function without resorting to Lagrangian duality techniques, hence adding more flexibility in the choice of the regularization function. The obtained results demonstrate the advantages of using nonsmooth sparsity-inducing regularization in this context. More specifically, we have shown that the $\ell_{1,\infty}$-norm constitutes a good choice for preventing overfitting in the case when just a few training examples are available. Furthermore, our experiments indicate that the epigraphical method converges much faster than the solution based on standard techniques for handling linear constraints.

## 6. REFERENCES


[1] C. Cortes and V. Vapnik, "Support-vector networks," *J. Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sept. 1995.

[2] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–392, 2001.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 23-28 Jun. 2008, pp. 1–8.

[4] D. Martín-Iglesias, J. Bernal-Chaves, C. Peláez-Moreno, A. Gallardo-Antolín, and F. Díaz-de María, "A speech recognizer based on multiclass SVMs with HMM-guided segmentation," *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 3817, pp. 257–266, 2005.

[5] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, 2005.

[6] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electronic Computers*, vol. EC-14, no. 3, pp. 326–334, June 1965.

[7] M. Aizerman, E. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[8] T. Joachims, T. Finley, and C.-N. J. Yu, "Cutting-plane training of structural SVMs," *J. Mach. Learn.*, vol. 77, no. 1, pp. 27–59, Oct. 2009.

[9] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations Trends in Mach. Learn.*, vol. 4, no. 1, pp. 1–106, 2012.

[10] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan, "A dirty model for multi-task learning," in *Adv. Neural Inf. Process. Syst. 23*, pp. 964–972. 2010.

[11] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, "An efficient projection for $\ell_{1,\infty}$ regularization," in *Int. Conf. Mach. Learn.*, Montreal, Quebec, Jun., 14-18 2009.

[12] L. Wang, X. Shen, and Y. F. Zheng, "On $L_1$-norm multi-class support vector machines," in *ICMLA*, Orlando, USA, 14-16 Dec. 2006, pp. 83–88.

[13] J. Langford, L. Li, and T. Zhang, "Sparse online learning via truncated gradient," *J. Mach. Learn.*, vol. 10, pp. 777–801, 2009.

[14] R. I. Boţ, A. Heinrich, and G. Wanka, "Employing different loss functions for the classification of images via supervised learning," 2013, preprint, www.mat.univie.ac.at/ rabot/publications/jour13-13.pdf.

[15] M. Blondel, K. Seki, and K. Uehara, "Block coordinate descent algorithms for large-scale sparse multiclass classification," *J. Mach. Learn.*, vol. 93, no. 1, pp. 31–52, Oct. 2013.

[16] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale $\ell_1$-regularized logistic regression," *J. Mach. Learn. Res.*, vol. 8, pp. 1519–1555, 2007.

[17] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Trans. Pattern Anal. Match. Int.*, vol. 27, no. 6, June 2005.

[18] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, New York, 2011.

[19] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame-based inverse problems," *Inverse Problems*, vol. 23, no. 4, Jun. 2007.

[20] P. L. Combettes and J.-C. Pesquet, "A proximal decomposition method for solving convex variational inverse problems," *Inverse Problems*, vol. 24, no. 6, Dec. 2008.

[21] J. M. Fadili and G. Peyré, "Total variation projection with first order schemes," *IEEE Trans. Img. Proc.*, vol. 20, no. 3, pp. 657–669, Mar. 2011.

[22] B. Taskar, C. Guestrin, and D. Koller, "Max-margin Markov networks," *Advances in Neural Information Processing Systems*, vol. 16, pp. 25–32, 2004.

[23] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, Springer-Verlag, 2004.

[24] G. Chierchia, N. Pustelnik, J.-C. Pesquet, and B. Pesquet-Popescu, "Epigraphical splitting for solving constrained convex formulations of inverse problems with proximal tools," 2013, Submitted, http://arxiv.org/pdf/1210.5844.pdf.

[25] J.-C. Pesquet and N. Pustelnik, "A parallel inertial proximal optimization method," *Pac. J. Optim.*, vol. 8, no. 2, pp. 273–305, Apr. 2012.

[26] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imag. Vis.*, vol. 40, no. 1, 2011.

[27] P. L. Combettes and J.-C. Pesquet, "Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators," *Set-Valued Var. Anal.*, vol. 20, no. 2, pp. 307–330, June 2012.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[29] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Trans. Pattern Anal. Match. Int.*, vol. 35, no. 8, pp. 1872–1886, 2013.