Learning SVMs from Sloppily Labeled Data

Guillaume Stempfel and Liva Ralaivola

Laboratoire d'Informatique Fondamentale de Marseille
Aix-Marseille Université
{guillaume.stempfel,liva.ralaivola}@lif.univ-mrs.fr

Abstract. This paper proposes a modelling of Support Vector Machine (SVM) learning to address the problem of learning with *sloppy labels*. In binary classification, learning with sloppy labels is the situation where a learner is provided with labelled data, where the observed labels of each class are possibly noisy (flipped) version of their true class and where the probability of flipping a label y to $-y$ only depends on y. The noise probability is therefore constant and uniform within each class: learning with *positive and unlabeled data* is for instance a motivating example for this model. In order to learn with sloppy labels, we propose SloppySvm, an SVM algorithm that minimizes a tailored nonconvex functional that is shown to be a uniform estimate of the noise-free SVM functional. Several experiments validate the soundness of our approach.

1 Introduction

This paper addresses the problem of learning a Support Vector Machine (SVM) from sloppily labelled data, that is from labelled data where the observed labels are possibly noisy versions of the true labels. We focus on binary classification and the noise process that we consider is the one where the probability of flipping a label to the opposite class depends only on the true label; the flipping probabilities (for the +1 and -1 classes) are therefore constant for each class.

Beyond simple theoretical motivations, being able to learn a large margin classifier is of interest for a few practical situations. If the training data at hand are manually labelled, there are reasons to believe that the labelling process is not perfect and that mislabelling may occur, and uniform mislabelling may be a seen as a consequence of the tiresomeness of the manual annotating task. Even more interesting is the connection between learning with sloppy labels and semi-supervised learning where the (few) labelled data available come from only one class; this is a common situation in, e.g., bioinformatics. This connection is formalized in Section 2 where it is also discussed how learning with sloppily labelled data may help learn in the multi-instance framework.

In order to tackle the problem of learning an SVM from sloppy labels, we propose to minimize a new objective functional, that is shown to be a *uniform estimate* of the noise-free SVM objective functional. The minimization of this nonconvex functional is done using a classical quasi-Newton minimization algorithm, where the nonconvex functional is rendered differentiable using the

smoothing trick proposed in [3] for the hinge loss. We also propose a heuristic to automatically estimate the noise levels from the training data.

The paper is organized as follows. Section 2 describes the learning setting and shows how it is related to the problem of semi-supervised learning with positive and unlabeled data only. Section 3 describes the new nonconvex objective functional that we propose to minimize together with our heuristic to automatically estimate the noise levels. Finally, Section 4 reports a number of numerical simulations that support the soundness of our proposed learning method.

2 Formal Setting and Motivations

2.1 Notation

Focusing on binary classification, the target space is $\mathcal{Y} = \{-1, +1\}$ and the input space $\mathcal{X}$ is assumed to be a *Hilbert space* with inner product $\langle \cdot, \cdot \rangle$. The family of classifiers we consider is that of zero-bias hyperplanes on $\mathcal{X}$ defined as $\langle \mathbf{w}, \mathbf{x} \rangle = 0$.[1] The class predicted for $\mathbf{x}$ using $\mathbf{w}$ is given by $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$. Even if we describe our results in the context of linear classifiers, they naturally carry over to the case of kernel classifiers.

Given a fixed (unknown) distribution D on $\mathcal{X} \times \mathcal{Y}$, a noise-free sample $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is made of n data independently and identically distributed according to D. Throughout, we assume: $\exists R > 0 : \mathbb{P}_{\mathbf{x} \sim D}(\langle \mathbf{x}, \mathbf{x} \rangle \leq R^2) = 1$.

Let us introduce $\mathcal{N} = \{\boldsymbol{\eta} : \boldsymbol{\eta} = [\eta^+ \, \eta^-] \in [0, 1)^2, \eta^+ + \eta^- < 1\}$. The sloppy labelling is defined with respect to a noise vector $\boldsymbol{\eta} = [\eta^+, \eta^-] \in \mathcal{N}$ such that a sample $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is corrupted by independently flipping each label y_i to $-y_i$ with probability η^{y_i}, where $\eta^y = \eta^+$ if $y = +1$ and $\eta^y = \eta^-$ otherwise. This noise process can be modeled using a Rademacher vector $\boldsymbol{\sigma} = [\sigma_1 \cdots \sigma_n]$ of size n, with $\mathbb{P}(\sigma_i = -1) = \eta^{y_i} = 1 - \mathbb{P}(\sigma_i = 1)$, to give the noisy version $\mathcal{S}^\sigma = \{(\mathbf{x}_i, \sigma_i y_i)\}_{i=1}^n$ of $\mathcal{S}$. In this paper, we will assume that the noise vector $\boldsymbol{\eta} \in \mathcal{N}$ *is known* (and fixed). We discuss in the conclusion how a reliable estimate of this vector can be carried out from the data.

The *margin* $\gamma : \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is defined by $\gamma(\mathbf{w}, \mathbf{x}, y) = y \langle \mathbf{w}, \mathbf{x} \rangle$. $\ell : \mathbb{R} \to \mathbb{R}^+$ denotes the hinge loss: $\ell(\gamma) = \max(0, 1 - \gamma)$.

Tackled problem We address the problem of learning a large margin separating hyperplane from $\mathcal{S}^{\boldsymbol{\sigma}} = \{(\mathbf{x}_i, \sigma_i y_i)\}_{i=1}^n$, where large margin must be understood with respect to the noise-free sample $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

2.2 Main Motivation: Asymmetric Semi-Supervised Learning

Our primary motivation to study the possibility of learning a large margin classifier from sloppily labelled data is that of being able to perform semi-supervised

[1] The inclusion of a bias term is straightforward and would not change our message. For sake of clarity and conciseness, we choose to consider zero-bias hyperplanes only.

learning when the labelled data only come from one class, which is called *asymmetric semi-supervised learning* [4]. In this situation, the training data can be written as $\mathcal{S} = \mathcal{S}_{+1} \cup \mathcal{S}_{\text{unl}}$ where $\mathcal{S}_{+1} = \{(\mathbf{x}_i, 1)\}_{i=1}^{m}$ and $\mathcal{S}_{\text{unl}} = \{\mathbf{x}_i\}_{i=m+1}^{n}$, and where the data from $\mathcal{S}_{+1}$ are supposed to be actual (i.e. noise free) positive data. The strategy that we envision to tackle this problem is to arbitrarily label all the data from $\mathcal{S}_{\text{unl}}$ as negative data, to give $\mathcal{S}_{-1} = \{(\mathbf{x}_i, -1)\}_{i=m+1}^{n}$, which in turn gives rise to a new training set $\underline{\mathcal{S}} = \mathcal{S}_{+1} \cup \mathcal{S}_{-1}$. Hence, as $\mathcal{S}_{\text{unl}}$ possibly contained positive instances, the data of $\underline{\mathcal{S}}$ are sloppily labelled data with $\eta^+ > 0$ (the positive data from $\mathcal{S}_{\text{unl}}$ are erroneously labelled as -1) while $\eta^- = 0$ (no negative data is labelled as $+1$). This way of tackling the problem of learning from positive and unlabeled data only was successfully undertaken in [4].

Note that the problem of multi-instance learning can also be cast, to some extent, in a problem of learning with sloppy labels. Indeed, in multi-instance learning, the training set is of the form $\mathcal{S} = \{(\underline{\mathbf{x}}_i, y_i)\}$ where each $\underline{\mathbf{x}}_i$ is a *bag of descriptions*, i.e. $\underline{\mathbf{x}}_i = \{\mathbf{x}_i^1, \ldots, \mathbf{x}_i^{n_i}\}$ where $\mathbf{x}_i^j \in \mathcal{X}$. A bag is classified as $+1$ if at least one element of the bag is indeed positive. A strategy to address the multi-instance learning problem is simply to break each multi-instance labelled pair $(\underline{\mathbf{x}}_i, y_i)$ in n_i labelled pairs $(\mathbf{x}_i^j, y_i)$ and to consider the question of learning from $\underline{\mathcal{S}} = \{(\mathbf{x}_i^j, y_i)\}_{i,j=1}^{n,n_i}$. Hence, the negative instances from $\underline{\mathcal{S}}$ are indeed negative, whereas some of the positive instances are in fact negative (recall that it suffices for a bag to contain one positive element for it to be positive). If each bag of instances roughly contains the same ratio of positive instances, then the problem is that of learning from the sloppy training set $\underline{\mathcal{S}}$ with $\eta^+ = 0$ and $\eta^- > 0$.

This paper intends to provide a *first step* to envision tackling asymmetric semi-supervised learning with SVMs.

3 Proposed Approach

3.1 SloppySVM: a Version of CSVM Robust to Sloppy Labels

Recall that the CSVM problem for a sample $\mathcal{S}$ of size n writes as (see, e.g., [6])

$$\min_{\mathbf{w},b} \frac{1}{2}\langle \mathbf{w}, \mathbf{w} \rangle + \frac{C}{n} \sum_{i=1}^{n} \ell(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle). \tag{1}$$

where $C \in \mathbb{R}^+$ is a regularization parameter. It is clear that when a sloppily labelled dataset $\mathcal{S}^{\boldsymbol{\sigma}}$ is fed to the learning algorithm there is no reason for the $\mathbf{w}$ minimizing (1) to be a valid classifier because of the evaluation of the slack/hinge errors, accounted for by the second term of (1). Therefore, if it were possible to estimate the value of the noise-free slack errors, it would be possible to accurately learn a large margin classifier from the noisy data. In this section, we actually show that such an estimation is possible.

Let $\boldsymbol{\eta} = [\eta^+ \, \eta^-] \in \mathcal{N}$ be fixed and let K_η denote $K_\eta = \frac{1}{1-\eta^+-\eta^-}$ from now on. Let us introduce the mapping $\hat{\ell} : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, with

$$\hat{\ell}(\gamma, y) = K_\eta \left[(1 - \eta^{-y}) \ell(\gamma) - \eta^y \ell(-\gamma) \right]. \tag{2}$$

The following lemma holds.

Lemma 1. $\forall i \in \{1,\dots,n\} : \mathbb{E}_{\sigma_i}\hat{\ell}(\sigma_i y_i\langle \mathbf{w},\mathbf{x}_i\rangle, \sigma_i y_i) = \ell(y_i\langle \mathbf{w},\mathbf{x}_i\rangle)$. *Therefore,*

$$\mathbb{E}_{\boldsymbol{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\hat{\ell}(\sigma_i y_i\langle \mathbf{w},\mathbf{x}_i\rangle, \sigma_i y_i) = \frac{1}{n}\sum_{i=1}^{n}\ell(y_i\langle \mathbf{w},\mathbf{x}_i\rangle).$$

Proof. The proof is straightforward. Assume that y_i is fixed (hence is the distribution of σ_i). Let us introduce $\gamma_i = y_i\langle \mathbf{w},\mathbf{x}_i\rangle$. We have

$$\begin{aligned}
\mathbb{E}_{\sigma_i}\hat{\ell}(\sigma_i\gamma_i, \sigma_i y_i) &= (1-\eta^{y_i})\hat{\ell}(\gamma_i, y_i) + \eta^{y_i}\hat{\ell}(-\gamma_i, -y_i)\\
&= K_\eta\left[(1-\eta^{y_i})\left[(1-\eta^{-y_i})\ell(\gamma_i) - \eta^{y_i}\ell(-\gamma_i)\right] + \eta^{y_i}\left[(1-\eta^{y_i})\ell(-\gamma_i) - \eta^{-y_i}\ell(\gamma_i)\right]\right]\\
&= K_\eta\left[(1-\eta^{y_i})(1-\eta^{-y_i})\ell(\gamma_i) - \eta^{y_i}\eta^{-y_i}\ell(\gamma_i)\right]\\
&= K_\eta(1-\eta^{y_i}-\eta^{-y_i})\ell(\gamma_i) = \ell(\gamma_i).
\end{aligned}$$

The linearity of the expectation gives the second result of the lemma. □

This lemma says that, for given parameter $\mathbf{w}$, we can estimate the noise-free slack/fitting errors from the noisy data. Using $\hat{\ell}$ we therefore propose a new version of CSvm based on noisy data. Given a sloppy dataset $\mathcal{S}^{\boldsymbol{\sigma}} = \{(\mathbf{x}_i, \sigma_i y_i)\}_{i=1}^n$, this new learning strategy, called SloppySvm, aims at solving

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + \frac{C}{n}\sum_{i=1}^{n}\hat{\ell}(\sigma_i y_i\langle \mathbf{w},\mathbf{x}_i\rangle, \sigma_i y_i). \tag{3}$$

First observe that the objective function of SloppySvm can actually be computed from the sloppy dataset $\mathcal{S}^{\boldsymbol{\sigma}}$. Additionally, note that (a) if the noise rates are such that $\eta^+ = 0$ and $\eta^- = 0$ then problem (3) boils down to the classical CSvm problem (1), (b) the expectation of the objective function of (3) with respect to the noise process is the objective function of (1). Finally, notice that even though the objective function of (1) is convex, that of (3) is not necessarily convex, because of the nonconvexity of $\hat{\ell}$ (see (2)). Figure 1 illustrates the behaviors of the mappings $\hat{\ell}_y$, for $y = 1$ and $y = -1$, defined as:

$$\hat{\ell}_y : \mathbb{R} \to \mathbb{R}, \quad \gamma \mapsto \hat{\ell}_y(\gamma) = \hat{\ell}(\gamma, y). \tag{4}$$

3.2 Uniform Closeness of the New Functional to its Expectation

The question we address now is to quantify how close the objective of (3) is to the objective of the corresponding noise free CSvm problem (for the same value of C and the same instances $\mathbf{x}_i$), i.e. its expectation. We show using concentration inequalities and properties on the Rademacher complexity of (kernel) linear classifier that uniform bounds (with respect to $\mathbf{w}$) on the closeness of these functionals can be drawn.

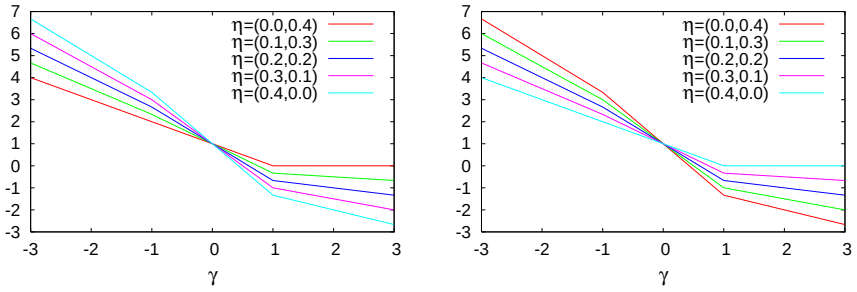

Fig. 1: Left: $\hat{\ell}_{+1}$ (resp. Right: $\hat{\ell}_{-1}$) as a function of γ for different $\boldsymbol{\eta} = [\eta^+ \, \eta^-]$ (see (2)); the $\hat{\ell}_y$'s are nonconvex and Lipschitz with constant $K_\eta(1 + \left|\eta^+ - \eta^-\right|)$ (cf. Lemma 2).

To do so, we only consider vectors $\mathbf{w}$ such that $\|\mathbf{w}\| < W$, for some $W > 0$. We are therefore to investigate the closeness of

$$\mu(\mathcal{S}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \text{ and } \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \hat{\ell}(\sigma_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle, \sigma_i y_i). \quad (5)$$

In order to state our main proposition (Proposition 1), we first introduce and/or recall some results.

Lemma 2. *Let L_η be the constant defined by $L_\eta = 1 - |\eta^+ - \eta^-|$. The mappings $\hat{\ell}_y$, for $y = +1$ and $y = -1$ are Lipschitz with constant $K_\eta L_\eta$.*

Proof. It suffices to observe that the slopes of $\gamma \mapsto \hat{\ell}_y(\gamma)$ are (see also Figure 1), $K_\eta(1 - \eta^{-y})$ if $\gamma < -1$, $K_\eta(1 - \eta^{-y} + \eta^y)$ if $-1 \leq \gamma < 1$, and $K_\eta \eta^y$ otherwise. Observing that $\eta^y \leq 1 - \eta^{-y} + \eta^y$ and $1 - \eta^{-y} < 1 - \eta^{-y} + \eta^y$ together with $1 - \eta^{-y} + \eta^y < 1 + |\eta^+ - \eta^-| = L_\eta$ ends the proof. □

Theorem 1 (McDiarmid [5]). *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set $\mathcal{X}$, and assume that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies*

$$\sup_{\substack{\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X} \\ \mathbf{x}_i' \in \mathcal{X}}} |f(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n) - f(\mathbf{x}_1, \ldots, \mathbf{x}_i', \ldots, \mathbf{x}_n)| \leq c_i$$

for every $1 \leq i \leq n$. Then, for every $t > 0$,

$$\mathbb{P}\left\{|f(X_1, \ldots, X_n) - \mathbb{E}f(X_1, \ldots, X_n)| \geq t\right\} \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

Theorem 2 (Bartlett and Mendelson [1]). *Let F be a class of real functions, and let*

$$R_n(F) = \mathbb{E}_{\mathcal{S}\boldsymbol{\kappa}} \frac{2}{n} \sup_{f \in F} \left| \sum_{i=1}^{n} \kappa_i f(\mathbf{x}_i) \right|$$

be the Rademacher complexity *of F for samples of size n (κ_i are i.i.d random Rademacher variables, i.e. $\mathbb{P}(\kappa_i = +1) = \mathbb{P}(\kappa_i = -1) = \frac{1}{2}$). The following holds:*

– *If* $\varphi : \mathbb{R} \to \mathbb{R}$ *is* L_φ*-Lipschitz and* $\varphi(0) = 0$*, then* $R_n(\varphi \circ F) \leq 2L_\varphi R_n(F)$
– *If* $F = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle, \|\mathbf{w}\| \leq W\}$ *then* $R_n(F) \leq \frac{2WR}{\sqrt{n}}$.

Note that given a distribution D on $\mathcal{X} \times \mathcal{Y}$, the noise process that gives rise to sloppy labels entails a distribution on $\mathcal{X} \times \mathcal{Y} \times \{-1, 1\}$ for triples of the form (X, Y, σ), with $\mathbb{P}(\sigma = 1|Y = y) = 1 - \eta^y$; conditioning on $Y\sigma = t$ for $t = \{-1, +1\}$, we may therefore define the conditional distribution $D_{XY\sigma|Y\sigma=t}$ over $\mathcal{X} \times \mathcal{Y} \times \{-1, 1\}$. Then, given vector $\mathbf{t} = [t_1 \cdots t_n]$ with $t_i \in \{+1, -1\}$, $D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}} = \otimes_{i=1}^n D_{XY\sigma|Y\sigma=t_i}$ is the distribution of samples $\mathcal{S} = \{(X_i, Y_i)\}_{i=1}^n$ and noise vector $\boldsymbol{\sigma} = [\sigma_1 \cdots \sigma_n]$ such that $Y_i\sigma_i = t_i$. Note that all the triples (X_i, Y_i, σ_i) such that $t_i = 1$ ($t_i = -1$) share the same distribution – and are, of course, independent. The meaning of $D^n_{\mathcal{S}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}$ is easily deduced.

Proposition 1. *Let* $\mathbf{t} = [t_1 \dots t_n]$ *be a fixed vector of* $\{-1, +1\}^n$*. For all distributions* $D, \forall \boldsymbol{\eta} = [\eta^+ \eta^-] \in \mathcal{N}$*,* $\forall \delta \in (0, 1]$*,* $\forall \varepsilon \in \mathbb{R}^+$*, for all random sample* $\mathcal{S}$ *and noise vector* $\boldsymbol{\sigma}$ *of size* n *drawn from* $D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}$ *if*

$$n \geq \max\left(\frac{8K_\eta^2 L_\eta^2 (1 + RW)^2}{\varepsilon^2} \ln \frac{4}{\delta}, \frac{256\,(2K_\eta L_\eta WR + 1)^2}{\varepsilon^2}\right)$$

then, with probability at least $1 - \delta$*,*

$$|\mu(\mathcal{S}, \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})| < \varepsilon, \quad \forall \mathbf{w} \in \mathcal{X}, \|\mathbf{w}\| \leq W. \tag{6}$$

Proof. Here, expectations must be understood with respect to $D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}$ and $D^n_{\mathcal{S}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}$, n^+ (n^-) is the number t_i equal to 1 (-1). We derive a uniform (wrt $\mathbf{w}$, $\|\mathbf{w}\| \leq W$) bound on $|\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\mu(\mathcal{S}, \mathbf{w}) - \mu(\mathcal{S}, \mathbf{w})|$ and on $|\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})|$ (cf equation (5)). It turns out that an adequate sample size for the latter to be lower than $\varepsilon > 0$ is sufficient for the former to be lower than ε as well (the proof omitted for sake of conciseness but it can be easily deduced from the present proof). We thus focus on bounding the function Δ defined as

$$\Delta(\mathcal{S}, \boldsymbol{\sigma}) = \sup_{\mathbf{w} \in \mathcal{X}, \|\mathbf{w}\| \leq W} \left|\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}}\hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})\right|.$$

Since $\|\mathbf{w}\| \leq W$ and $\|\mathbf{x}\| \leq R$, the minimum and maximum achievable margin by $\mathbf{w}$ on any pair $(\mathbf{x}, y)$ are $\gamma_{\min} = -RW$ and $\gamma_{\max} = RW$, respectively. Hence, since $\hat{\ell}_y$ is a decreasing function of γ and according to (4), $\hat{\ell}_y$ takes values in the range (both for $y = -1$ and $y = +1$) $[\hat{\ell}_y(\gamma_{\max}); \hat{\ell}_y(\gamma_{\min})]$, which is a subset of

$$\left[-K_\eta \max(\eta^+, \eta^-)(1 + RW); K_\eta(1 - \min(\eta^+, \eta^-))(1 + RW)\right].$$

Therefore, using again $\gamma_i = y_i\langle \mathbf{w}, \mathbf{x}_i \rangle$, the maximum variation of $\hat{\ell}_{t_i}(\sigma_i\gamma_i)$ when changing any triple $(\mathbf{x}_i, y_i, \sigma_i)$ to another one is at most

$$K_\eta(1 + RW)\left[\max(\eta^+, \eta^-) + \max(1 - \eta^+, 1 - \eta^-)\right] = K_\eta L_\eta (1 + RW).$$

Hence, given the definition of $\hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})$ (cf. (5)), the maximum variation of $\Delta(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})$ when changing any $(\mathbf{x}_i, y_i, \sigma_i)$ is at most $K_\eta L_\eta (1 + RW)/n$. Using Theorem 1, we thus have (the triples (X_i, Y_i, σ_i) are independent of each other)

$$\mathbb{P}\left\{ \left| \Delta(\mathcal{S}, \boldsymbol{\sigma}) - \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \Delta(\mathcal{S}, \boldsymbol{\sigma}) \right| \geq \frac{\varepsilon}{4} \right\} \leq 2 \exp\left(- \frac{n\varepsilon^2}{8 K_\eta^2 L_\eta^2 (1 + RW)^2} \right),$$

which is upper bounded by $\delta/2$ for the choice of n stated in the lemma.

Then, we have the following upper bounding on $\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \Delta(\mathcal{S}, \boldsymbol{\sigma})$ (the sup is taken over $\mathbf{w} \in \mathcal{X}, \|\mathbf{w}\| \leq W$), where $\boldsymbol{\kappa}$ is a vector of n independent Rademacher variables and $\mathcal{S}'\boldsymbol{\sigma}' \sim D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}$, and, thus, $\mathcal{S}' \sim D^n_{\mathcal{S}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}$:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \Delta(\mathcal{S}, \boldsymbol{\sigma}) &= \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \sup \left| \mathbb{E}_{\mathcal{S}'\boldsymbol{\sigma}'} \hat{\mu}(\mathcal{S}', \boldsymbol{\sigma}', \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) \right| \\
&\leq \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \sup \mathbb{E}_{\mathcal{S}'\boldsymbol{\sigma}'} \left| s\hat{\mu}(\mathcal{S}', \boldsymbol{\sigma}', \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) \right| && \text{(triangle ineq.)} \\
&\leq \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}\mathcal{S}'\boldsymbol{\sigma}'} \sup \left| \hat{\mu}(\mathcal{S}', \boldsymbol{\sigma}', \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) \right| && \text{(Jensen ineq.)} \\
&= \frac{1}{n} \mathbb{E}_{\mathcal{S}\mathcal{S}'} \sup \left| \sum_{i=1}^n \hat{\ell}_{t_i}(t_i \langle \mathbf{w}, \mathbf{x}'_i \rangle) - \sum_{i=1}^n \hat{\ell}_{t_i}(t_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right| && \text{((5) and } \mathcal{S}, \mathcal{S}' \sim D^n_{\mathcal{S}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}) \\
&= \frac{1}{n} \mathbb{E}_{\mathcal{S}\mathcal{S}'\boldsymbol{\kappa}} \sup \left| \sum_{i=1}^n \kappa_i \left(\hat{\ell}_{t_i}(t_i \langle \mathbf{w}, \mathbf{x}'_i \rangle) - \hat{\ell}_{t_i}(t_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right) \right| && (\mathbf{x}_i \text{ and } \mathbf{x}'_i \text{ are i.i.d}) \\
&\leq \frac{2}{n} \mathbb{E}_{\mathcal{S}\boldsymbol{\kappa}} \sup \left| \sum_{i=1}^n \kappa_i \hat{\ell}_{t_i}(t_i \langle \mathbf{w}, \mathbf{x}_i \rangle) \right| && \text{(triangle ineq.)} \\
&\leq \frac{2}{n} \left[\mathbb{E}_{\mathcal{S}^+\boldsymbol{\kappa}} \sup \left| \sum_{i:t_i=+1} \kappa_i \hat{\ell}_{+1}(\langle \mathbf{w}, \mathbf{x}_i \rangle) \right| + \mathbb{E}_{\mathcal{S}^-\boldsymbol{\kappa}} \sup \left| \sum_{i:t_i=-1} \kappa_i \hat{\ell}_{-1}(-\langle \mathbf{w}, \mathbf{x}_i \rangle) \right| \right],
\end{aligned}$$

where $\mathcal{S}^+ \sim D^{n^+}_{\mathcal{S}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=[1\ldots1]}$ and $\mathcal{S}^- \sim D^{n^-}_{\mathcal{S}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=[-1\ldots-1]}$.

To further upper bound $\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \Delta(\mathcal{S}, \boldsymbol{\sigma})$, we introduce $Q_{n^+}(W)$ defined as

$$Q_{n^+}(W) = \frac{2}{n^+} \mathbb{E}_{\mathcal{S}^+\boldsymbol{\kappa}} \sup_{\mathbf{w}, \|\mathbf{w}\| \leq W} \left| \sum_{i=1}^{n^+} \kappa_i \hat{\ell}_{+1}(\langle \mathbf{w}, \mathbf{x}_i \rangle) \right|,$$

and we observe that

$$\begin{aligned}
Q_{n^+}(W) &\leq \frac{2}{n^+} \mathbb{E}_{\mathcal{S}^+\boldsymbol{\kappa}} \sup \left| \sum_{i=1}^{n^+} \kappa_i \hat{\ell}_{+1}(\langle \mathbf{w}, \mathbf{x}_i \rangle) - 1 \right| + \frac{2}{n^+} \mathbb{E}_{\boldsymbol{\kappa}} \left| \sum_{i=1}^{n^+} \kappa_i \right| && \text{(triangle ineq.)} \\
&\leq \frac{2}{n^+} \mathbb{E}_{\mathcal{S}^+\boldsymbol{\kappa}} \sup \left| \sum_{i=1}^{n^+} \kappa_i \hat{\ell}_{+1}(\langle \mathbf{w}, \mathbf{x}_i \rangle) - 1 \right| + \frac{2}{\sqrt{n^+}},
\end{aligned}$$

since $\mathbb{E}_{\boldsymbol{\kappa}} \left| \sum_{i=1}^n \kappa_i \right|^2 = \mathbb{E}_{\boldsymbol{\kappa}} \left[\sum_{i,j=1}^n \kappa_i \kappa_j \right] = n$ and for any random variable X, $0 \leq \mathbb{V}(|X|) = \mathbb{E}(|X|^2) - \mathbb{E}^2(|X|)$ and, thus, $\mathbb{E}(|X|) \leq \sqrt{\mathbb{E}(|X|^2)}$.

From Lemma 2, $\hat{\ell}_{+1}(\gamma) - 1$ is Lipschitz with constant $K_\eta L_\eta$ and $\hat{\ell}_{+1}(0) - 1 = 0$. The first term of the last inequality is the Rademacher complexity of the

class of functions defined by the composition of $\hat{\ell}() - 1$ and the set of zero-bias hyperplanes $\mathbf{w}$ such that $\|\mathbf{w}\| \leq W$. Using Theorem 2, we therefore have:

$$Q_{n^+}(W) \leq \frac{4K_\eta L_\eta WR}{\sqrt{n^+}} + \frac{2}{\sqrt{n^+}} = \frac{2(2K_\eta L_\eta WR + 1)}{\sqrt{n^+}},$$

while a similar inequality holds for the counterpart $Q_{n^-}(W)$ of $Q_{n^+}(W)$. Hence,

$$\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \Delta(\mathcal{S}, \boldsymbol{\sigma}) \leq \frac{n^+}{n} \frac{2(2K_\eta L_\eta WR + 1)}{\sqrt{n^+}} + \frac{n^-}{n} \frac{2(2K_\eta L_\eta WR + 1)}{\sqrt{n^-}} \leq \frac{4(2K_\eta L_\eta WR + 1)}{\sqrt{n}}$$

which, for the value of n stated in the lemma is upper bounded by $\varepsilon/4$. Therefore, with probability at least $1-\delta/2$ the following holds uniformly over $\mathbf{w}$, $\|\mathbf{w}\| \leq W$:

$$|\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})| \leq \frac{\varepsilon}{4} + \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \Delta(\mathcal{S}, \boldsymbol{\sigma}) \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

Likewise, $|\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \mu(\mathcal{S}, \mathbf{w}) - \mu(\mathcal{S}, \mathbf{w})| \leq \varepsilon/2$ with probability $1-\delta/2$ as well. Noting that $\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \mu(\mathcal{S}, \mathbf{w}) = \mathbb{E}_{\mathcal{S}} \mu(\mathcal{S}, \mathbf{w}) = \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})$ and using

$$|\mu(\mathcal{S}, \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})| \leq |\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \mu(\mathcal{S}, \mathbf{w}) - \mu(\mathcal{S}, \mathbf{w})| + |\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma}} \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})|$$

ends the proof. □

Proposition 2. *With the same hypotheses as in Proposition 1, for $\mathcal{S}$ drawn according to D (no conditioning on $\mathbf{Y}\boldsymbol{\sigma}$), and corresponding random vector $\boldsymbol{\sigma}$, the following holds with probability at least $1-\delta$:*

$$|\mu(\mathcal{S}, \mathbf{w}) - \hat{\mu}(\mathcal{S}, \boldsymbol{\sigma}, \mathbf{w})| < \varepsilon, \quad \forall \mathbf{w} \in \mathcal{X}, \|\mathbf{w}\| \leq W. \tag{7}$$

Proof. It Φ denotes the event given by equation (6), then we just stated that $\forall \mathbf{t} \in \{-1, +1\}^n$, $\mathbb{P}_{\mathbf{S}\boldsymbol{\sigma} \sim D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}}(\Phi) \geq 1-\delta$. Then,

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}\sim D}(\Phi) &= \mathbb{E}_{\mathbf{T}}[\mathbb{E}_{\mathcal{S}\boldsymbol{\sigma} \sim D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{T}}} \mathbb{I}_\Phi] = \sum_{\mathbf{t}} \mathbb{E}_{\mathcal{S}\boldsymbol{\sigma} \sim D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}} \mathbb{I}_\Phi \mathbb{P}(\underline{\mathbf{Y}\boldsymbol{\sigma}} = \mathbf{t}) \\
&= \sum_{\mathbf{t}} \mathbb{P}_{\mathcal{S}\boldsymbol{\sigma} \sim D^n_{\mathcal{S}\boldsymbol{\sigma}|\underline{\mathbf{Y}\boldsymbol{\sigma}}=\mathbf{t}}}(\Phi) \mathbb{P}(\mathbf{Y}\boldsymbol{\sigma} = \mathbf{t}) \geq \sum_{\mathbf{t}} (1-\delta) \mathbb{P}(\mathbf{Y}\boldsymbol{\sigma} = \mathbf{t}) = 1 - \delta.
\end{aligned}$$

□

This establishes the closeness of the functional of SloppySvm to the original Svm functional and justifies the strategy that consists in minimizing (3), when sloppily labeled data are available.

3.3 A Quasi-Newton Optimization Method

In order to actually minimize (3), we use the BFGS quasi-Newton minimization procedure directly applied to the primal. As proposed by [3], we used a twice-differentiable approximation L_h of the hinge loss function ℓ, with, for $h > 0$

$$L_h(\gamma) := \begin{cases} 0 & \text{if } \gamma > 1 + h \\ \frac{(1+h-\gamma)^2}{4h} & \text{if } |1 - \gamma| \leq h \\ 1 - \gamma & \text{if } \gamma < 1 - h \end{cases} \tag{8}$$

Plugging in this loss function in (3) as a surrogate for ℓ, we end up with an unconstrained minimization problemfor which it is easy to find a local minimum.

4 Numerical Simulations

In order to illustrate the behavior of SloppySvm on real data, we have carried out simulations on 5 Uci datasets, preprocessed and made available by Gunnar Rätsch[2]. These numerical simulations consist in adding a class dependent noise to the data, and then learn on the sloppy datasets with CSvm, and SloppySvm, the actual noise levels being provided to SloppySvm. For each problem, the data are split into 100 training and test sets (except for Splice, where only 20 replicates are available). All the experiments have been made using a linear kernel, as they make it possible in the noise free case to achieve nearly 'optimal' accuracies. Several values (from 0.005 to 1000) of the parameter C have been tested. For each problem and for each noise level, we picked the value of C that minimizes the generalization error of CSvm. For SloppySvm, whichever the level of noise, we use the value of C that minimizes the generalization error when CSvm is trained on the noise-free data.

The results of the simulations are reported in Table 1, which shows the error rates and the standard deviations computed on the replicates of the different problems. It is clear that, even if the behavior of CSvm may seem satisfactory on some problems, especially when η^+ and η^- are close to each other, CSvm is consistently outperformed by SloppySvm on all the conducted experiments. The performances of SloppySvm are close, even under rather high noise levels, to the performance obtained by CSvm on noise-free datasets, while keeping a relatively low standard deviation. This shows the stability of SloppySvm.

5 Conclusion

We have proposed SloppySvm, a new large margin algorithm that can deal with sloppily labeled data. It makes use of a new objective function that computes an estimator the noise-free fitting term of usual CSvm. A theoretical analysis ensures the closeness of the proposed objective functional to the noise-free one.

[2] http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm

$[\eta^+, \eta^-]$	*Alg.*	*Diabetis*	*Flare-Solar*	*German*	*Heart*	*Splice*
$[0, 0]$	All	23.3 ± 1.59	32.32 ± 1.81	24.03 ± 2.24	15.31 ± 3.17	16.11 ± 0.56
$[0, 0.2]$	CSvm	25.18 ± 2.08	38.99 ± 6.15	25.32 ± 2.65	18.11 ± 4.02	19.20 ± 1.02
	Sloppy	*23.69 ± 1.74*	*32.45 ± 1.73*	*24.67 ± 2.04*	*16.19 ± 3.78*	*17.10 ± 0.70*
$[0, 0.4]$	CSvm	44.06 ± 11.9	44.74 ± 1.81	44.77 ± 7.81	30.10 ± 6.48	29.64 ± 2.14
	Sloppy	*24.02 ± 1.79*	*33.91 ± 4.08*	*26.06 ± 2.47*	*17.68 ± 3.88*	*18.71 ± 0.68*
$[0.2, 0]$	CSvm	25.00 ± 2.37	37.45 ± 3.02	27.93 ± 3.12	18.78 ± 3.9	18.24 ± 1.10
	Sloppy	*23.63 ± 1.64*	*32.78 ± 1.66*	*24.40 ± 2.31*	*16.63 ± 4.01*	*16.93 ± 0.10*
$[0.2, 0.4]$	CSvm	33.83 ± 9.93	44.57 ± 2.52	37.39 ± 6.39	29.97 ± 8.10	28.08 ± 2.17
	Sloppy	*26.04 ± 2.96*	*35.65 ± 4.32*	*28.32 ± 3.14*	*20.73 ± 5.27*	*23.67 ± 1.73*
$[0.4, 0]$	CSvm	33.93 ± 2.93	41.43 ± 2.89	29.82 ± 1.95	27.97 ± 6.42	28.3 ± 2.14
	Sloppy	*24.1 ± 1.91*	*33.37 ± 2.71*	*25.36 ± 2.4*	*17.18 ± 3.75*	*18.08 ± 0.90*
$[0.4, 0.2]$	CSvm	33.49 ± 3.56	40.61 ± 3.5	29.82 ± 1.95	26.96 ± 7.06	27.84 ± 2.42
	Sloppy	*26.29 ± 2.80*	*35.85 ± 4.15*	*28.80 ± 2.96*	*21.66 ± 7.08*	*23.33 ± 2.08*

Table 1: Generalization errors (and standard deviations) obtained from simulations conducted on 5 datasets from the UCI repository when corrupted by various amount of noise. Italicized results indicate best errors (not necessary statistically significant).

Using a standard quasi-Newton minimization procedure, numerical simulations conducted on noisy problems prove the soundness of our approach.

We have carried out preliminary experiments on asymmetric semi-supervised learning problems where the noise is estimated from the data. The procedure of noise estimation is that proposed in [2] for co-training, which is proved to be consistent: among candidate noise vectors, this suggests to select the noise $\boldsymbol{\eta}$ that minimizes the quantity $\mathbb{P}(h(x) = 1|y\sigma = -1) + \mathbb{P}(h(x) = -1|y\sigma = 1)$, where h is a learned SloppySvm using $\boldsymbol{\eta}$. The results we obtain are very promising and we plan to investigate this empirical evaluation more thoroughly.

References

1. P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. of Machine Learning Research*, 3:463–482, 2002.
2. A. Blum and T. Mitchell. Combining Labeled and Unlabeled Data with Co-Training. In *Proc. of the 11th Conf. on Computational Learning Theory*, pages 92–100, 1998.
3. O. Chapelle. Training a support vector machine in the primal. *Neural Comput.*, 19(5):1155–1178, 2007.
4. C. Magnan. Asymmetrical Semi-Supervised Learning and Prediction of Disulfide Connectivity in Proteins. *RIA, New Methods in Machine Learning: Theory and applications*, 20(6):673–695, 2006.
5. C. McDiarmid. On the method of bounded differences. *Survey in Combinatorics*, pages 148–188, 1989.
6. B. Schölkopf and A. J. Smola. *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. MIT University Press, 2002.