LINEAR SUPPORT VECTOR MACHINES WITH NORMALIZATIONS

Yiyong Feng and Daniel P. Palomar

Dept. of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong

ABSTRACT

In this paper, we start with the standard support vector machine (SVM) formulation and extend it by proposing a general SVM that allows many different variations captured by normalizations in the formulation with very diverse numerical performance. The proposed formulation can not only capture the existing work, i.e., standard soft-margin SVM, ℓ_1-SVM, as special cases, but also enable us to propose more SVMs that outperform the existing ones under some scenarios.

Index Terms— Convex Optimization, Normalizations, SVM.

1. INTRODUCTION

Since the support vector machine (SVM) was established [1, 2], it has become the standard technique for many different supervised classification problems in different fields, e.g., the cancer diagnostic in bioinformatics, image classification in objective detection, face recognition in computer vision, text categorization in document processing, and for more related applications, see [3].

The standard soft-margin SVM usually leads to nonsparse solutions. However, in many real applications it is imperative to perform feature selection to detect which features are actually relevant. The common way of doing it is with a sparsity penalty [4]. Some examples are the classic ℓ_1-norm penalty [5], the exponential concave penalty [6], or adding convex relaxation constraints on ℓ_1-norm penalty [7].

Feature scaling can be treated as a generalization of feature selection by weighting or scaling the features with different scalars rather than only 0 or 1. From this point of view, the method of standardizing the input data before training the SVM is a special case of feature scaling where each feature is independently normalized so that it has zero mean and unit variance. However, all the knowledge of the location and scale of the original data may be lost after standardization [8, 9] and there is no guarantee that standardization will improve the classification performance in general [10]. Still, data standardization is useful to avoid the features with larger dynamic range dominating those with smaller ones and the numerical difficulties during the calculation [8, 9]. Some methods have been proposed to find better feature scaling. Papers [11] and [12] focused on finding the optimal feature scaling via minimizing some analytical upper bounds on the leave-one-out cross validation error, since the gradient of such objectives with respect to the scaling variables can be easily computed and the simple gradient method can be implied to find at least some local optimal solution easily. Later, an adaptive method [13] was proposed to avoid the potential overfitting.

However, the aforementioned methods all lead to nonconvex problems. To overcome this drawback, [14] proposed the concept "normalized margin" and the optimization problem of maximizing "normalized margin" can be reformulated into a convex form revealing a connection with the traditional ℓ_1-SVM; in particular as a weighted ℓ_1-SVM where the weights can be computed based on the input data directly.

In this paper, motivated by the work in [14], we propose a general SVM formulation that can capture many existing linear SVMs as special cases by taking different normalizations, show the connections and differences between different SVMs clearly, and moreover provide us with more insights on different SVMs. We also benefit from the proposed formulation by proposing some new SVMs that outperform the existing ones under small training size scenarios. This proposed formulation is the main contribution of our work.

This work was supported by the Hong Kong RGC 617312 research grant.

Notation We adopt the notation of using boldface lower case for vectors $\mathbf{a}$, upper case for matrices $\mathbf{A}$. The notation $\mathbf{1}$ denotes all one column vector with proper size. The transpose operator is $(\cdot)^\top$, the trace operator is $\mathrm{Tr}(\cdot)$, and $\|\cdot\|_F$ means the matrix Frobenius norm. The curled inequality symbol $\succeq$ is used to denote generalized inequality: $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is an Hermitian positive semidefinite matrix. The element of matrix $\mathbf{A}$ at the i-th row and j-th column is denoted by $\mathbf{A}_{ij}$. The notation $\mathbf{A}^{1/2}$ denotes principal square root of matrix $\mathbf{A}$. $\mathrm{Diag}(\mathbf{A})$ denotes a diagonal matrix with diagonal elements being that of $\mathbf{A}$, and its principal square root is $\mathrm{Diag}^{1/2}(\mathbf{A})$. The notation $\mathcal{R}(\mathbf{X})$ stands for the range space of $\mathbf{X}$.

2. LINEAR SUPPORT VECTOR MACHINES

For a binary classification problem: $\mathbf{x}_i \in \mathbb{R}^d \to y_i \in \{+1, -1\}$, $i = 1, 2, \ldots, N$, the goal of linear classification is to find a linear decision boundary that classifies $\mathbf{x}_i$'s according to their labels y_i.

The soft-margin SVM aiming to find the trade-off between the large margin and small misclassification has the formulation [2]:

$$\begin{array}{ll} \underset{\beta_0,\boldsymbol{\beta},\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top\mathbf{x}_i + \beta_0\right) \geq 1 - \xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{1}$$

It turns out that $1/\|\boldsymbol{\beta}\|_2$ has a nice interpretation that it measures the separation between the two classes. For example, for the linear separable case, it equals to the minimum distance between the samples from either class and the linear decision boundary. Because of that, the quantity is also called "margin".

A well-known method to induce feature selection or sparsity in $\boldsymbol{\beta}$ consists in replacing the ℓ_2-norm with ℓ_1-norm[1] [5]:

$$\begin{array}{ll} \underset{\beta_0,\boldsymbol{\beta},\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{\beta}\|_1^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top\mathbf{x}_i + \beta_0\right) \geq 1 - \xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{2}$$

3. A GENERAL LINEAR SVM FORMULATION

3.1. Problem statement

Consider the general linear mapping:

$$\varphi(\mathbf{x}_i, \mathbf{F}) \triangleq \mathbf{F}\mathbf{x}_i \tag{3}$$

[1] Usually $\|\boldsymbol{\beta}\|_1$ is used in the objective rather than $\|\boldsymbol{\beta}\|_1^2$. However, those two formulations are equivalent for an appropriate choice of C. For the consistency of presentation in this paper, we adopt $\|\boldsymbol{\beta}\|_1^2$ here.

where $\mathbf{F} \in \mathbb{R}^{m\times d}$. If $\mathbf{F}$ is square and diagonal, the mapping is scaling the features so that they are independent of the units in which they were measured [14]. The more general nonsquare and nondiagonal $\mathbf{F}$ allows for more degrees of freedom; for example, the features can be rotated prior to the scaling. Here, we extend the concept of normalized margin (NM) proposed in [14] by taking more general "normalization" as follows:

$$\mathrm{NM} \triangleq M/\sqrt{\phi(\mathbf{F})} \tag{4}$$

where M is the margin, and $\phi(\mathbf{F})$ is a general normalization term that measures how compact the training data is. Rather than focusing on some specific definition of the normalization $\phi(\mathbf{F})$, we suppose a general assumption as stated below.

Assumption 1. *We assume $\phi(\mathbf{F})$ is function of $\mathbf{F}$ and can always be written in the following form*

$$\phi(\mathbf{F}) \triangleq \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}}^{1/2}\right\|_F^2 \tag{5}$$

where matrix $\mathbf{A}_{\boldsymbol{\theta}} \in \mathbb{R}^{d\times d}$ is positive semi-definite and represents the information abstracted from the training data $\{\mathbf{x}_i, y_i\}_{i=1}^N$ indexed by the parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Note that $\phi(\mathbf{F})$ defined by (5) is convex in $\mathbf{F}$ since it is the pointwise maximum of a family of quadratic convex functions of $\mathbf{F}$ indexed by $\boldsymbol{\theta}$. To understand how $\phi(\mathbf{F})$ measures the compactness of the training data, let us visit the normalization term used in [14]:

Example 1. *The normalization in [14]*

$$\phi(\mathbf{F}) \triangleq \sum_{i,j=1}^{N} \frac{1+y_iy_j}{2} \left\|\mathbf{F}(\mathbf{x}_i-\mathbf{x}_j)\right\|_2^2 \tag{6}$$

uses the summation of squared distances among the same class data instances to measure the compactness of the transformed training samples, and it is a specific example of (5) with

$$\boldsymbol{\Theta} \triangleq \{0\}, \tag{7}$$

$$\mathbf{A}_0 \triangleq \sum_{i,j=1}^{N} \frac{1+y_iy_j}{2} (\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top. \tag{8}$$

To keep the problem statement as general as possible, we make use of the general expression (5) for the normalization $\phi(\mathbf{F})$ in the following part of this section. The underlying idea is to maximize the margin while making each class as compact as possible. To start with, we consider the linearly separable case first. The problem of jointly finding the linear mapping $\mathbf{F}$ and the parameters of the separating hyperplane with the normalized margin maximized can be formulated as:

$$\begin{array}{ll} \underset{M,\boldsymbol{\beta},\beta_0,\mathbf{F}}{\text{maximize}} & M/\sqrt{\phi(\mathbf{F})} \\ \text{subject to} & y_i\frac{1}{\|\boldsymbol{\beta}\|_2}\left(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0\right) \geq M, \quad \forall i. \end{array} \tag{9}$$

Since for any feasible $\boldsymbol{\beta}$ and β_0, any positively scaled multiple is also feasible and we can arbitrarily set $\|\boldsymbol{\beta}\|_2 = 1/M$ and problem (9) can be reformulated as:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\beta_0,\mathbf{F}}{\text{minimize}} & \frac{1}{2}\phi(\mathbf{F})\|\boldsymbol{\beta}\|_2^2 \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0\right) \geq 1, \quad \forall i. \end{array} \tag{10}$$

Similar to the soft-margin SVM (1), the linearly nonseparable case of maximizing normalized margin problem can be formulated as:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\beta_0,\mathbf{F},\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\phi(\mathbf{F})\|\boldsymbol{\beta}\|_2^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{11}$$

The idea of the linear transformation $\mathbf{F}\mathbf{x}$ here is quite similar to many problems in signal processing, for example, the optimal linear precoding designs for the MIMO communication systems [15] or wideband noncooperative systems [16].

3.2. Proposed solving approach

Due to space limitations, we provide our theoretical findings only in this paper. The detailed proofs, more meaningful explorations and insights, and more numerical experiments are presented in a full version of this work [17].

Obviously, the problem of interest (11) is nonconvex. Fortunately, we are able to reformulate it into convex form. Since $\phi(\mathbf{F}) = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}}\right)\right\}$ is quadratic in $\mathbf{F}$ for any given $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, we can always scale $\mathbf{F}$ and $\boldsymbol{\beta}$ appropriately so that $\phi(\mathbf{F}) = 1$. Furthermore, the constraint can be relaxed to $\phi(\mathbf{F}) \leq 1$ since the equality will always be active at the optimal point, otherwise we could scale $\boldsymbol{\beta}$ down and scale $\mathbf{F}$ up with the same scalar to find another feasible point but with the objective value further reduced. Then, problem (11) is equivalent to:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\beta_0,\mathbf{F},\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \phi(\mathbf{F}) \leq 1, \\ & \boldsymbol{\xi} \geq \mathbf{0}, \end{array} \tag{12}$$

which can be further reformulated as:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\beta_0,\mathbf{F},\boldsymbol{\xi},t}{\text{minimize}} & \frac{1}{2}t + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \boldsymbol{\beta}\boldsymbol{\beta}^\top \preceq t\mathbf{I}, \\ & \phi(\mathbf{F}) \leq 1, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{13}$$

To proceed, we consider the following different problem:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\beta_0,\mathbf{F},\boldsymbol{\xi},t}{\text{minimize}} & \frac{1}{2}t + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \mathbf{F}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{F} \preceq t\mathbf{F}^\top\mathbf{F}, \\ & \phi(\mathbf{F}) \leq 1, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{14}$$

Interestingly, we have the following result.

Proposition 1. *Problem (13) and problem (14) have the same optimal value and their optimal solutions have the following relationships:*

- *If $(\boldsymbol{\beta}_1^\star, \beta_{01}^\star, \mathbf{F}_1^\star, \boldsymbol{\xi}_1^\star, t_1^\star)$ is a optimal solution of problem (13), then it is also a optimal solution of problem (14);*
- *If $(\boldsymbol{\beta}_2^\star, \beta_{02}^\star, \mathbf{F}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is a optimal solution of problem (14), then $\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star, \beta_{02}^\star, \mathbf{F}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star\right)$ is a optimal solution of problem (13), where $\mathbf{P}_{\mathcal{R}\left(\mathbf{F}_2^\star\right)}$[2] is the projector that projects any vector onto $\mathcal{R}(\mathbf{F}_2^\star)$.*

[2] $\mathbf{P}_{\mathcal{R}(\mathbf{X})} = \mathbf{X}\mathbf{X}^\dagger$, where $\mathbf{X}^\dagger$ is the pseudo inverse of $\mathbf{X}$ [18].

In other words, Prop. 1 simply says that problem (13) and problem (14) are equivalent, and thus we can investigate problem (14) instead. Denote the variables $\mathbf{v} \triangleq \mathbf{F}^\top \boldsymbol{\beta}$, $\mathbf{T} \triangleq \mathbf{F}^\top \mathbf{F} \succeq \mathbf{0}$ ($\mathbf{T} \in \mathbb{R}^{d\times d}$), and by the Schur complement, problem (14) can be rewritten as a semi-definite programming (SDP):

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\mathbf{T},\boldsymbol{\xi},t}{\text{minimize}} & \frac{1}{2}t + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\mathbf{v}^\top\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\mathrm{Tr}\left(\mathbf{T}\mathbf{A}_{\boldsymbol{\theta}}\right)\right\} \leq 1, \\ & \left[\begin{array}{cc} t & \mathbf{v}^\top \\ \mathbf{v} & \mathbf{T} \end{array}\right] \succeq \mathbf{0}, \\ & \mathrm{rank}\left(\mathbf{T}\right) \leq \min\left(m,d\right), \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{15}$$

Note that, if $m < d$ the above problem is still nonconvex due to the the rank constraint $\mathrm{rank}\left(\mathbf{T}\right) \leq m$. However, under some condition, we can show that the SDP relaxation (SDR) in fact is tight.

Proposition 2. *Problem (15) is bounded below by:*

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\left\|\mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}\right\|_2^2\right\} + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\mathbf{v}^\top\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{16}$$

In addition, if there exists some $\mathbf{A}_{\boldsymbol{\theta}}$ full rank, the lower bound is tight, and the optimal solution of (16) is the optimal solution of (15) with $\mathbf{T} = \mathbf{v}\mathbf{v}^\top / \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\left\|\mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}\right\|_2^2\right\}$.

Thus, when there exists some $\mathbf{A}_{\boldsymbol{\theta}}$ full rank, problem (11) and (16) are indeed equivalent no matter the size of $\mathbf{F}$.

Next, we revisit the case $\mathbf{F}$ being diagonal. Similar to the derivation procedure from (11) to (15), we can easily check that problem (11) can be reformulated as:

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\mathbf{T},\boldsymbol{\xi},t}{\text{minimize}} & \frac{1}{2}t + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\mathbf{v}^\top\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\mathrm{Tr}\left(\mathbf{T}\mathbf{A}_{\boldsymbol{\theta}}\right)\right\} \leq 1, \\ & \left[\begin{array}{cc} t & \mathbf{v}^\top \\ \mathbf{v} & \mathbf{T} \end{array}\right] \succeq \mathbf{0}, \\ & \mathbf{T}\,\text{is diagonal}, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{17}$$

Furthermore, we can have the following result.

Proposition 3. *Problem (17) is bounded below by:*

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\left\|\mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\theta}}\right)\mathbf{v}\right\|_1^2\right\} + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\mathbf{v}^\top\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{18}$$

In addition, if there exists some $\mathrm{Diag}\left(\mathbf{A}_{\boldsymbol{\theta}}\right)$ full rank, the lower bound is tight, and the optimal solution of (18) is the optimal solution of (17) with $\mathbf{T}_{ii} = |\mathbf{v}_i| / \left(\sqrt{\mathbf{A}_{\boldsymbol{\theta}^\star ii}}\left\|\mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\theta}^\star}\right)\mathbf{v}\right\|_1\right)$ where $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\left\|\mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\theta}}\right)\mathbf{v}\right\|_1^2\right\}$.

In fact, problem (23) in [14] is a specific case of problem (18) such that $\boldsymbol{\Theta}$ and $\mathbf{A}_{\boldsymbol{\theta}}$ are given by (7) and (8). However, our proof in [17] is much simpler and more straightforward.

3.3. Classification prediction

Once the parameters $(\mathbf{v}, \beta_0, \boldsymbol{\xi})$ are trained from the SVMs (e.g., (16) and (18)), we can have the separating hyperplane: $f\left(\mathbf{x}\right) \triangleq \mathbf{v}^\top\mathbf{x}+\beta_0$ and the classification prediction for some new outcome sample $\mathbf{x}$ simply is $\hat{y} = \mathrm{sign}\left(f\left(\mathbf{x}\right)\right)$.

3.4. Insights and new SVMs

For sake of clarity, we allow ourselves a slight abuse of notation $\mathbf{A}_{\boldsymbol{\theta}}$ from case to case in the following part of this section.

If we compare Props. 2 and 3, we can see that different normalizations result in different penalties:

- When only scaling is considered, i.e., $\mathbf{F}$ is restricted to be a square diagonal matrix in mapping (3), the normalized margin can be interpreted as the reciprocal of the weighted ℓ_1-norm of the normal vector of the separating hyperplane.
- When $\mathbf{F}$ is allowed to be any m-by-d matrix in mapping (3), the normalized margin can be interpreted as the reciprocal of the weighted ℓ_2-norm of the normal vector of the separating hyperplane.

Revisiting the soft-margin SVM (1) and ℓ_1-SVM (2), we can see they both do not consider the information in the training data since $\mathbf{A}_{\boldsymbol{\theta}} = \mathbf{I}$ (they still differ from each other by adopting different vector norms). Then the previous interpretations might indicate that the SVMs (16) and (18) based on normalized margin formulation may be able to improve the classification performance for some data sets since they can incorporate data structure information via the weigh matrices $\mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\theta}}\right)$ or $\mathbf{A}_{\boldsymbol{\theta}}$ into the problem formulations.

Inspired by the the above observation, we want to propose more specific SVMs with data structure information incorporated.

3.4.1. Measuring compaction within each class

Consider $\boldsymbol{\Theta}$ and $\mathbf{A}_0$ are given by (7) and (8), and we use the penalty

$$\begin{aligned} \left\|\left(\mathbf{A}_0+\nu\mathbf{I}\right)^{1/2}\mathbf{v}\right\|_2^2 &= \sum_{i,j=1}^N \frac{1+y_iy_j}{2}\left(f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right)^2 + \nu\mathbf{v}^\top\mathbf{v} \\ &= \sum_{y_i,y_j=1}\left(f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right)^2 + \sum_{y_i,y_j=-1}\left(f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right)^2 + \nu\mathbf{v}^\top\mathbf{v} \end{aligned} \tag{19}$$

where $\nu \geq 0$ is a trade-off parameter. Since $f\left(\mathbf{x}_i\right)$ represents the signed distance of point $\mathbf{x}_i$ to the separating hyperplane up to a common positive scalar[3], thus the first two terms of (19) represents the summation of squared differences among the distances of the samples to the separating hyperplane within each class. Then penalizing (19) means finding the trade-off between concrete compaction of the samples (e.g., small $\sum_{i,j=1}^N \frac{1+y_iy_j}{2}\left(f\left(\mathbf{x}_i\right)-f\left(\mathbf{x}_j\right)\right)^2$) and large soft-margin (e.g., small $\mathbf{v}^\top\mathbf{v}$, since $\frac{1}{\sqrt{\mathbf{v}^\top\mathbf{v}}}$ stands for the soft-margin).

3.4.2. Measuring compaction within the whole training data set

Similarly, we can also select $\mathbf{A}_0 = \sum_{i=1}^N\left(\mathbf{x}_i-\bar{\mathbf{x}}_0\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}_0\right)^\top$, where $\bar{\mathbf{x}}_0 \triangleq \frac{1}{N}\sum_{i=1}^N\mathbf{x}_i$, and use the following penalty instead

$$\left\|\left(\mathbf{A}_0+\nu\mathbf{I}\right)^{1/2}\mathbf{v}\right\|_2^2 = \sum_{i,j=1}^N\left(f\left(\mathbf{x}_i\right)-f\left(\bar{\mathbf{x}}_0\right)\right)^2 + \nu\mathbf{v}^\top\mathbf{v}. \tag{20}$$

[3] Actually, $\left(\mathbf{v}^\top\mathbf{x}_i+\beta_0\right) / \left\|\mathbf{v}\right\|_2$ is exactly the signed distance from point $\mathbf{x}_i$ to the hyperplane $\mathbf{v}^\top\mathbf{x}+\beta_0 = 0$, see [3, Eq. (4.40)].

Table 1. The simulated SVMs: existing I-III, proposed IV-V. Here $[\mathbf{X}]_{F=d} \triangleq \sqrt{d}\frac{\mathbf{X}}{\|\mathbf{X}\|_F}$ so that $\left\|[\mathbf{X}]_{F=d}\right\|_F = \sqrt{d}$.

(UF)
$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\left\|\mathbf{A}_0^{1/2}\mathbf{v}\right\|_p^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\mathbf{v}^\top\mathbf{x}_i+\beta_0\right) \geq 1-\xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array}$$

No.	p	$\mathbf{A}_0$
I	1	$\left[\text{Diag}\left(\sum_{i,j=1}^N \frac{1+y_iy_j}{2}\left(\mathbf{x}_i-\mathbf{x}_j\right)\left(\mathbf{x}_i-\mathbf{x}_j\right)^\top\right)\right]_{F=d}$
II	2	$\mathbf{I}$
III	1	$\mathbf{I}$
IV	2	$\left[\sum_{i,j=1}^N \frac{1+y_iy_j}{2}\left(\mathbf{x}_i-\mathbf{x}_j\right)\left(\mathbf{x}_i-\mathbf{x}_j\right)^\top\right]_{F=d} + \nu\mathbf{I}$
V	2	$\left[\sum_{i=1}^N \left(\mathbf{x}_i-\bar{\mathbf{x}}_0\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}_0\right)^\top\right]_{F=d} + \nu\mathbf{I}$

where the first term, e.g., the summation of squared differences among the distances between the samples to the separating hyperplane and the distance between the sample mean to the separating hyperplane, measures how compact the data is.

The underlying ideas of the above two penalty examples, e.g., (19) and (20), are really the same: the samples within each class should be somehow compact with respect to the separating hyperplane. The only difference is that they use different quantities to measure the compactness of the data.

Intuitively, the above two examples make sense, especially when the number of training samples is small. Because when we do not have large enough number of training samples, simply maximizing the soft-margin, e.g., penalizing $\mathbf{v}^\top\mathbf{v}$ only, may give us the separating hyperplane that has wrong direction and could be really misleading. However, once we have taken the data information into the consideration of the formulation, e.g., penalizing (19) or (20), we may get better separating hyperplanes that are closer to the true one.

4. NUMERICAL EXPERIMENTS

In this section, we compare the proposed methods, e.g., IV-V in Table 1, and the existing ones, e.g., I-III in Table 1. All the optimization problems are solved via the commercial solver MOSEK [19] in MATLAB. Due to space limitations, we leave the detailed numerical setup in the full version [17].

To illustrate the insights in Section 3.4 clearly, we consider a visualizable synthetic experiment with only two attributes, that is, $d = 2$. Here we consider two classes with equal probabilities. Samples of class $+1$ are drawn from $\mathcal{N}(\mathbf{1}_d, \boldsymbol{\Sigma})$ and samples of class -1 are drawn from $\mathcal{N}(-\mathbf{1}_d, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the common covariance matrix $\boldsymbol{\Sigma} = 4 \times \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$.

Fig. 1a shows the results when number of training samples is small. The shape of the distributions is shown by the green ellipses as contours and the optimal boundary is the solid red line. We can see that the existing SVM II aiming to maximize the soft-margin may be misleading sometimes (see the blue dotted line) since only a few support vectors matter and the other samples do not affect the separating hyperplane at all. Fortunately, the proposed methods take all the training data into account and thus aim at finding the large margin and let all the samples within the same class be compact to each other can provide separating hyperplanes closer to the optimal boundary (see the black dash-dotted line and magenta dashed line). Fig. 1b shows the Receiver Operating Characteristic (ROC) curves of different methods. We can clearly see that the two proposed methods outperform the standard soft-margin SVM.

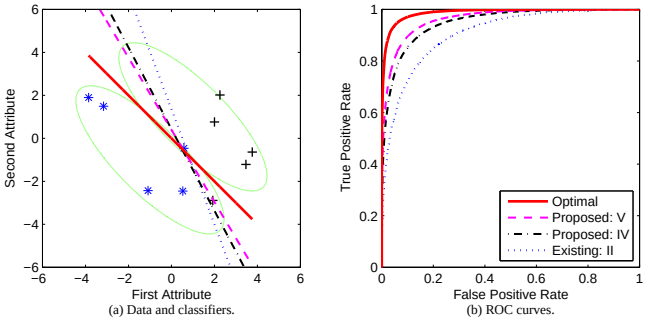

Fig. 1. Synthetic example when number of training samples is small.

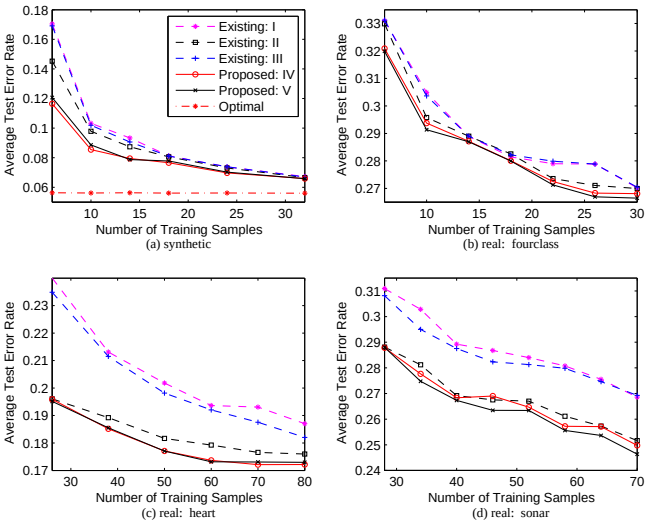

Fig. 2. Average test error rate versus number of training samples.

Fig. 2a shows best average test error rate versus the number of training samples of the synthetic data. Clearly, we can see that the proposed methods are comparable or even outperform the existing ones, especially when the training sample size is small. This verifies the insights behind Fig. 1.

Figs. 2b-d show the results of three real data sets available from [20] (no theoretical optimal boundary anymore). We can observe similar results and in fact the improvements by the proposed methods look even more significant than that of the synthetic data.

5. CONCLUSION

In this paper, we have proposed a general linear SVM formulation that can characterize both the proposed and many existing SVMs by simply selecting different types of weighted vector norms. The origin of having such different methods to incorporate the general linear SVM problem formulation with different normalizations. The formulation can provide us with more insights and help us understand the connections and differences between different SVMs. The numerical experiments on both synthetic and real-world data sets show that the proposed methods can outperform the existing ones when the number of training samples is not large.

6. REFERENCES

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

[2] V. Vapnik, *The Nature of Statistical Learning Theory*. springer, 2000.

[3] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, New York, 2009.

[4] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[5] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Advances in Neural Information Processing Systems*, 2004, pp. 49–56.

[6] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *the 15th International Conference on Machine Learning*, Madison, WI, USA, June 1998, pp. 82–90.

[7] A. B. Chan, V. Nuno, and G. R. Lanckriet, "Direct convex relaxations of sparse SVM," in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, USA, June 2007, pp. 145–153.

[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[9] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*. SIAM, 2007, vol. 20.

[10] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Tech. Rep., 2003.

[11] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Machine Learning*, vol. 46, no. 1-3, pp. 131–159, 2002.

[12] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Information Processing Systems*, 2000, pp. 668–674.

[13] Y. Grandvalet and S. Canu, "Adaptive scaling for feature selection in SVMs," in *Advances in Neural Information Processing Systems*, 2002, pp. 553–560.

[14] M. H. Nguyen and F. De la Torre, "Optimal feature selection for support vector machines," *Pattern Recognition*, vol. 43, no. 3, pp. 584–591, 2010.

[15] D. P. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Foundations and Trends in Communications and Information Theory*, vol. 3, no. 4, pp. 331–551, 2006.

[16] G. Scutari, D. P. Palomar, and S. Barbarossa, "Optimal linear precoding strategies for wideband noncooperative systems based on game theory–part I: Nash equilibria," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1230–1249, 2008.

[17] Y. Feng and D. P. Palomar, "Normalization of linear support vector machines," *Submitted to IEEE Trans. Signal Process.*, 2014.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge university press, 1990.

[19] MOSEK, "The mosek optimization toolbox for MATLAB manual," Tech. Rep., 2013. [Online]. Available: http://www.mosek.com

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.