Normalization of Linear Support Vector Machines

Yiyong Feng and Daniel P. Palomar, *Fellow, IEEE*

***Abstract*—In this paper, we start with the standard support vector machine (SVM) formulation and extend it by considering a general SVM formulation with normalized margin. This results in a unified convex framework that allows many different variations in the formulation with very diverse numerical performance. The proposed unified framework can capture the existing methods, i.e., standard soft-margin SVM, ℓ_1-SVM, and SVMs with standardization, feature selection, scaling, and many more SVMs, as special cases. Furthermore, our proposed framework can not only provide us with more insights on different SVMs from the "energy" and "penalty" point of views, which help us understand the connections and differences between them in a unified way, but also enable us to propose more SVMs that outperform the existing ones under some scenarios.**

***Index Terms*—Convex optimization, normalizations, support vector machines, unified framework.**

I. Introduction

SINCE the support vector machine (SVM) was established [1]–[4], 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 [5], [6].

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 [7]. Some examples are the classic ℓ_1-norm penalty [8], the exponential concave penalty [9], or adding convex relaxation constraints on ℓ_1-norm penalty [10]. An alternative way is to introduce one more 0-1 variable to do hard feature selection for each input feature [11]. Paper [12] provided numerical studies for some specific data sets.

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 [13], [14] and there is no guarantee that standardization will improve the classification performance in general [15]. 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 [13], [14]. Some methods have been proposed to find better feature scaling. Papers [16] and [17] 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, [18] proposed an adaptive method to avoid potential overfitting.

However, the aforementioned methods all lead to nonconvex problems. To overcome this drawback, [19] proposed the concept "normalized margin" and the optimization problem of maximizing "normalized margin" can be reformulated into convex form. They were able to link their problem with the traditional ℓ_1-SVM and pointed out their problem indeed is a weighted ℓ_1-SVM where the weights can be computed based on the input data directly. More recent related works on SVM can be found in [20] and references therein.

In this paper, motivated by the work in [19], we propose a unified framework that can capture many existing linear SVMs (including [19]) as special cases by taking different normalizations, show the connections and differences between different SVMs clearly, and moreover provide more insights on different SVMs from the "energy" and "penalty" points of view. We also benefit from the unified framework by having some new SVMs that are comparable with or even better than the existing ones under small training size scenarios.

This paper is organized as follows. Basic standard linear SVMs are reviewed in Section II. An extended general linear SVM with normalized margin, the main result, and the detailed contributions are provided in Section IV. Then Sections IV–VI show the detailed procedure of obtaining the main result. More specifically, Section IV provides a solving approach for the extended general linear SVM, Section V explores the general linear SVM with more normalizations, and Section VI summarizes the general linear SVM and explorations to propose a unified framework. At last, Section VII presents the numerical experiments, and Section VIII concludes the paper.

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 Moore-Penrose pseudo-inverse operator is $(\cdot)^\dagger$, 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 $\succeq$ denotes the generalized inequality: $\mathbf{A} \succeq \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is an Hermitian positive semidefinite matrix. The element of matrix $\mathbf{A}$

Manuscript received October 06, 2014; revised January 14, 2015 and April 03, 2015; accepted June 03, 2015. Date of publication June 09, 2015; date of current version August 06, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gustau CCamps-Valls. This work was supported by the Hong Kong RGC 617312 research grant.

The authors are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: yiyong@ust.hk; palomar@ust.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2015.2443730

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 equal to 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 $\mathbf{X}$. The notation $v^\star((\cdot))$ stands for the optimal value of problem $(\cdot)$.

II. Linear Support Vector Machines

Consider 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 a linear classification problem is to find a linear decision boundary that classifies the $\mathbf{x}_i$ according to their binary labels y_i. Generally speaking, a linear separating hyperplane can be expressed as

$$f(\mathbf{x}) \triangleq \mathbf{v}^\top \mathbf{x} + \beta_0 \tag{1}$$

where $\mathbf{v}$ and β_0 are the linear classifier parameters to learn from the training data set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, and the classification prediction for a new outcome sample $\mathbf{x}$ simply is $\hat{y} = \mathrm{sign}(f(\mathbf{x}))$.

The soft-margin SVM aiming at finding the trade-off between the large margin and small misclassification has the following problem formulation [4]:

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

It turns out that $1/\|\boldsymbol{\beta}\|_2$ has a nice interpretation that 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] [8]:

$$\begin{aligned} \underset{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{\beta}\|_1^2 + C\mathbf{1}^\top \boldsymbol{\xi} \\ \text{subject to} \quad & y_i(\boldsymbol{\beta}^\top \mathbf{x}_i + \beta_0) \geq 1 - \xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \tag{3}$$

III. A General Linear SVM Formulation With Normalized Margin

A. Problem Statement

Consider the general linear mapping:

$$\varphi(\mathbf{x}_i, \mathbf{F}) \triangleq \mathbf{F}\mathbf{x}_i \tag{4}$$

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 [19]. 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 [19] by taking more general "normalization" as follows:

$$\mathrm{NM} \triangleq M/\sqrt{\phi(\mathbf{F})} \tag{5}$$

[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 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

$$\phi(\mathbf{F}) \triangleq \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \mathrm{Tr}(\mathbf{F}^\top \mathbf{F} \mathbf{A}_{\boldsymbol{\theta}}) \right\} \tag{6}$$

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 (6) 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 [19]:

Example 1: The normalization in [19]

$$\phi(\mathbf{F}) \triangleq \sum_{i,j=1}^N \frac{1 + y_i y_j}{2} \|\mathbf{F}(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \tag{7}$$

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 (6) with

$$\boldsymbol{\Theta} \triangleq \{0\}, \tag{8}$$

$$\mathbf{A}_0 \triangleq \sum_{i,j=1}^N \frac{1 + y_i y_j}{2} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top. \tag{9}$$

To keep the problem statement as general as possible, we make use of the general expression (6) 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, consider the linearly separable case first, and 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:

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

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 (10) can be reformulated as:

$$\begin{aligned} \underset{\boldsymbol{\beta}, \beta_0, \mathbf{F}}{\text{minimize}} \quad & \frac{1}{2}\phi(\mathbf{F})\|\boldsymbol{\beta}\|_2^2 \\ \text{subject to} \quad & y_i(\boldsymbol{\beta}^\top \mathbf{F}\mathbf{x}_i + \beta_0) \geq 1, \quad \forall i. \end{aligned} \tag{11}$$

Similar to the soft-margin SVM (2), the linearly nonseparable case of maximizing normalized margin problem can be formulated as:

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

TABLE I
SUMMARY OF DIFFERENT PROBLEMS IN THE UNIFIED FRAMEWORK

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

Type	$\boldsymbol{\Theta}$	$\mathbf{A}_{\boldsymbol{\theta}}$
1	$\{0\}$	$\mathbf{A}_0\in\left\{\begin{array}{c}(9),(23),(29),(31),(53),(58),\\ \text{Diag}\left((9)\right),\text{Diag}\left((23)\right),\\ \text{Diag}\left((29)\right),\text{Diag}\left((31)\right)\end{array}\right\}$
2	$\{+1,-1\}$	$(\mathbf{A}_{+1},\mathbf{A}_{-1})\in\left\{\begin{array}{c}((25),(26)),((33),(34)),\\ (\text{Diag}\left((25)\right),\text{Diag}\left((26)\right)),\\ (\text{Diag}\left((33)\right),\text{Diag}\left((34)\right))\end{array}\right\}$
3	$\boldsymbol{\Lambda}_0$	$\mathbf{A}_{\boldsymbol{\lambda}_0}\in\{(39)\}$
4	$\boldsymbol{\Lambda}_{+1}\times\boldsymbol{\Lambda}_{-1}$	$\left(\mathbf{A}_{\boldsymbol{\lambda}_{+1}},\mathbf{A}_{\boldsymbol{\lambda}_{-1}}\right)\in\{((44),(45))\}$

The idea of the linear transformation $\mathbf{Fx}$ here is quite similar to many problems in signal processing, for example, the optimal linear precoding designs for the multiple-input-multiple-output (MIMO) communication systems [21] or wideband noncooperative systems [22], and or in financial engineering, e.g., the worst-case factor loading matrix in robust portfolio selection problems [23].

B. Main Result and Contributions

Note that the extended general linear SVM formulation with normalized margin (12) is nonconvex and it allows us to consider two extensions: i) more general transform $\mathbf{F}$, and ii) more general distance measurements as the normalizations for the normalized margin. Fortunately, in this paper, we are able to show that (12) equals to a convex unified framework (UF) in Table I, in which the norm is controlled by the transform $\mathbf{F}$ and weight matrix is controlled by distance measurement. In addition, the detailed contributions of the extended general linear SVM formulation with normalized margin (12) and its equivalent convex formulation can be summarized in several aspects as follows:

- The proposed unified framework is general enough to characterize many linear SVMs in a unified form via different normalizations or, more generally, different weighted vector norm penalties (as shown in Table I).
- The proposed unified framework enables us to understand different SVMs in a unified way: to connect different SVMs via the relationship between their normalizations.
- The proposed unified framework provides us with more meaningful insights: different normalizations/penalties mean different compactness measures among the data.
- The proposed unified framework benefits us with more new meaningful SVMs, e.g., the new SVMs we will explore later in Sections IV–VI.

In the following Sections IV–VI, we will propose an efficiently solving approach for (12), explore many more normalizations, gain some insights, and finally propose the convex unified framework (UF) and obtain the results in Table I. Meanwhile, we will also gain the above detailed contributions.

IV. PROPOSED SOLVING APPROACH

Obviously, the problem of interest (12) is nonconvex. Fortunately, we are able to reformulate it into a convex form. Since $\phi(\mathbf{F})=\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\text{Tr}(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}})\}$ 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 (12) is equivalent to:

$$\begin{array}{ll}\underset{\boldsymbol{\beta},\beta_0,\mathbf{F},\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2+C\mathbf{1}^\top\boldsymbol{\xi}\\ \text{subject to} & y_i(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0)\geq 1-\xi_i,\quad\forall i\\ & \phi(\mathbf{F})\leq 1,\\ & \boldsymbol{\xi}\geq\mathbf{0},\end{array}\tag{13}$$

which can be further reformulated as:

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

To proceed, we consider the following different problem:

$$\begin{array}{ll}\underset{\boldsymbol{\beta},\beta_0,\mathbf{F},\boldsymbol{\xi},t}{\text{minimize}} & \frac{1}{2}t+C\mathbf{1}^\top\boldsymbol{\xi}\\ \text{subject to} & y_i(\boldsymbol{\beta}^\top\mathbf{F}\mathbf{x}_i+\beta_0)\geq 1-\xi_i,\quad\forall i\\ & \mathbf{F}^\top\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{F}\preceq t\mathbf{F}^\top\mathbf{F},\\ & \phi(\mathbf{F})\leq 1,\\ & \boldsymbol{\xi}\geq\mathbf{0}.\end{array}\tag{15}$$

Interestingly, we have the following result.

Proposition 1: Problem (14) and problem (15) 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 (14), then it is also a optimal solution of problem (15);
- 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 (15), then $(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star,\beta_{02}^\star,\mathbf{F}_2^\star,\boldsymbol{\xi}_2^\star,t_2^\star)$ is a optimal solution of problem (14), where $\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}$[2] is the projector that projects any vector onto $\mathcal{R}(\mathbf{F}_2^\star)$.

Proof: See Appendix A. ∎

In other words, Prop. 1 simply says that problem (14) and problem (15) are equivalent, and thus we can investigate problem (15) 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 [25], problem (15) can be rewritten as a semi-definite programming (SDP) (recall that $\phi(\mathbf{F})=\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\text{Tr}(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}})\}$):

$$\begin{array}{ll}\underset{\mathbf{v},\beta_0,\mathbf{T},\boldsymbol{\xi},t}{\text{minimize}} & \frac{1}{2}t+C\mathbf{1}^\top\boldsymbol{\xi}\\ \text{subject to} & y_i(\mathbf{v}^\top\mathbf{x}_i+\beta_0)\geq 1-\xi_i,\quad\forall i\\ & \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\text{Tr}(\mathbf{T}\mathbf{A}_{\boldsymbol{\theta}})\}\leq 1,\\ & \begin{bmatrix}t & \mathbf{v}^\top\\ \mathbf{v} & \mathbf{T}\end{bmatrix}\succeq\mathbf{0},\\ & \text{rank}\,(\mathbf{T})\leq\min(m,d),\\ & \boldsymbol{\xi}\geq\mathbf{0}.\end{array}\tag{16}$$

[2] $\mathbf{P}_{\mathcal{R}(\mathbf{X})}=\mathbf{X}\mathbf{X}^\dagger$, where $\mathbf{X}^\dagger$ is the pseudo inverse of $\mathbf{X}$ [24].

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

Proposition 2: Problem (16) is bounded below by:

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\max\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\left\|\mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}\right\|_2^2\right\} + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i(\mathbf{v}^\top\mathbf{x}_i+\beta_0) \geq 1-\xi_i, \quad \forall i \\ & \boldsymbol{\xi}\geq\mathbf{0}. \end{array} \tag{17}$$

In addition, if there exists some $\mathbf{A}_{\boldsymbol{\theta}}$ full rank, the lower bound is tight, and the optimal solution of (17) is the optimal solution of (16) with $\mathbf{T} = \mathbf{v}\mathbf{v}^\top / \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\|\mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}\|_2^2\}$.

Proof: See Appendix B.

Thus, when there exists some $\mathbf{A}_{\boldsymbol{\theta}}$ full rank, problem (12) and (17) 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 (12)to (16), we can easily check that problem (12) can be reformulated as:

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\mathbf{T},\boldsymbol{\xi},t}{\text{minimize}} & \frac{1}{2}t + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i(\mathbf{v}^\top\mathbf{x}_i+\beta_0) \geq 1-\xi_i, \quad \forall i \\ & \max\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\mathrm{Tr}(\mathbf{T}\mathbf{A}_{\boldsymbol{\theta}})\} \leq 1, \\ & \begin{bmatrix} t & \mathbf{v}^\top \\ \mathbf{v} & \mathbf{T} \end{bmatrix} \succeq \mathbf{0}, \\ & \mathbf{T}\,\text{is diagonal}, \\ & \boldsymbol{\xi}\geq\mathbf{0}. \end{array} \tag{18}$$

Furthermore, we can have the following result.

Proposition 3: Problem (18) is bounded below by:

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\max\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\left\|\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})\mathbf{v}\right\|_1^2\right\} + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i(\mathbf{v}^\top\mathbf{x}_i+\beta_0) \geq 1-\xi_i, \quad \forall i \\ & \boldsymbol{\xi}\geq\mathbf{0}. \end{array} \tag{19}$$

In addition, if there exists some $\mathrm{Diag}(\mathbf{A}_{\boldsymbol{\theta}})$ full rank, the lower bound is tight, and the optimal solution of (19) is the optimal solution of (18) with $\mathbf{T}_{ii} = |\mathbf{v}_i|/(\sqrt{\mathbf{A}_{\boldsymbol{\theta}^\star ii}}\|\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}^\star})\mathbf{v}\|_1)$ where $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\|\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})\mathbf{v}\|_1^2\}$.

Proof: See Appendix C. ∎

In fact, problem (23) in [19] is a specific case of problem (19) such that $\boldsymbol{\Theta}$ and $\mathbf{A}_{\boldsymbol{\theta}}$ are given by (8) and (9). However, our proof in Appendix C is much simpler and more straightforward.

Remark 1: If we compare Props. 2 and 3, we can see that different normalizations/transformations result in different penalties:

- When only scaling is considered, i.e., $\mathbf{F}$ is restricted to be a square diagonal matrix in mapping (4), the normalized margin can be interpreted as the reciprocal of the weighted ℓ_1-norm of the normal vector of the separating hyperplane, where the weight matrix $\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})$ may represent the prior data structure information.
- When $\mathbf{F}$ is allowed to be any m-by-d matrix in mapping (4), the normalized margin can be interpreted as the reciprocal of the weighted ℓ_2-norm of the normal vector of the separating hyperplane, where the weight matrix $\mathbf{A}_{\boldsymbol{\theta}}$ may represent the prior data structure information.

The above interpretations might indicate that the SVMs (17) and (19) based on normalized margin formulation may be able to improve the classification performance for some data sets since they can consider data structure information in the problem formulations. This is an interesting observation and we will explore it in detail in this paper.

V. Exploration With More Normalizations

Recall that the underlying idea of normalized margin problem (12) is to maximize the margin and make the data to be compact at the same time. However, the aforementioned normalization (7) in [19] (e.g, Example 1) is only one specific example of the general normalization in Assumption 1. In this section, we will mainly explore the problem of interest (12) by considering more different normalizations.

Before proceeding, let us introduce several definitions. Without loss of generality, we assume that the first N_1 training samples have label $+1$ and the remaining $N - N_1$ training samples have label -1, i.e., $y_i = +1$ for $i = 1,\ldots,N_1$ and $y_i = -1$ for $i = N_1+1,\ldots,N$. Then we denote

$$\bar{\mathbf{x}}_0 \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i, \tag{20}$$

$$\bar{\mathbf{x}}_{+1} \triangleq \frac{1}{N_1}\sum_{i=1}^{N_1}\mathbf{x}_i, \quad \bar{\mathbf{x}}_{-1} \triangleq \frac{1}{N-N_1}\sum_{i=N_1+1}^{N}\mathbf{x}_i, \tag{21}$$

the means of the whole training samples, the samples within class $+1$, and the samples within class -1 respectively.

For sake of clarity, we allow ourselves the slight abuse of notations $\phi(\mathbf{F})$, $\boldsymbol{\Theta}$ and $\mathbf{A}_{\boldsymbol{\theta}}$ from case to case in the following part of this section.

A. Normalizations Based on Squared Distances Among Samples

1) Summation of Squared Distances Among Samples: The normalization is defined as the summation of squared distances among all the training samples:

$$\phi(\mathbf{F}) \triangleq \sum_{i,j=1}^{N}\|\mathbf{F}(\mathbf{x}_i-\mathbf{x}_j)\|_2^2 = \mathrm{Tr}(\mathbf{F}^\top\mathbf{F}\mathbf{A}_0) \tag{22}$$

where

$$\mathbf{A}_0 \triangleq \sum_{i,j=1}^{N}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top. \tag{23}$$

Remark 2: Note that the normalization (22) is a specific case of (6) with $\boldsymbol{\Theta} = \{0\}$ and $\mathbf{A}_0$ defined by (23). For the problem of interest (12) with this newly defined normalization (22), Prop. 2 and Prop. 3 apply for the cases of general $\mathbf{F}$ and diagonal $\mathbf{F}$ respectively.

2) Maximum of Squared Distances Among Samples Within Each Class: Instead of summation, we now consider the maximum of the squared distances among the training samples within the same classes as the normalization:

$$\begin{aligned}\phi(\mathbf{F}) &\triangleq \max\left\{\sum_{i,j=1}^{N_1}\|\mathbf{F}(\mathbf{x}_i-\mathbf{x}_j)\|_2^2, \sum_{i,j=N_1+1}^{N}\|\mathbf{F}(\mathbf{x}_i-\mathbf{x}_j)\|_2^2\right\} \\ &= \max\left\{\mathrm{Tr}(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{+1}), \mathrm{Tr}(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{-1})\right\}\end{aligned} \tag{24}$$

where

$$\mathbf{A}_{+1} \triangleq \sum_{i,j=1}^{N_1} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \tag{25}$$

$$\mathbf{A}_{-1} \triangleq \sum_{i,j=N_1+1}^{N} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top. \tag{26}$$

The matrix in (9) is the summation of (25) and (26).

Remark 3: Note that the normalization (24) is a specific case of (6) with $\boldsymbol{\Theta} = \{+1, -1\}$ and $\mathbf{A}_{+1}$ and $\mathbf{A}_{-1}$ defined by (25) and (26). For the problem of interest (12) with this newly defined normalization (24), Prop. 2 and Prop. 3 apply for the cases of general $\mathbf{F}$ and diagonal $\mathbf{F}$ respectively.

B. Normalizations Based on Squared Distances to the Center of Gravity

1) Summation of Squared Distances to the Center of Gravity: Now the normalization is defined as the summation of squared distances to the center of gravity:

$$\phi(\mathbf{F}) \triangleq \min_{\mathbf{c}} \sum_{i=1}^{N} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}\|_2^2. \tag{27}$$

It is not hard to see that $\mathbf{c}^\star = \frac{1}{N}\sum_{j=1}^{N} \mathbf{F}\mathbf{x}_j = \mathbf{F}\bar{\mathbf{x}}_0$, and thus

$$\phi(\mathbf{F}) = \sum_{i=1}^{N} \|\mathbf{F}\mathbf{x}_i - \mathbf{F}\bar{\mathbf{x}}_0\|_2^2 = \mathrm{Tr}(\mathbf{F}^\top \mathbf{F}\mathbf{A}_0) \tag{28}$$

where

$$\mathbf{A}_0 \triangleq \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}_0)(\mathbf{x}_i - \bar{\mathbf{x}}_0)^\top. \tag{29}$$

Remark 3: Note that the normalization (27), or equivalently (28), is a specific case of (6) with $\boldsymbol{\Theta} = \{0\}$ and $\mathbf{A}_0$ defined by (29). For the problem of interest (12) with this newly defined normalization (27), Prop. 2 and Prop. 3 apply for the cases of general $\mathbf{F}$ and diagonal $\mathbf{F}$ respectively.

2) Summation of Squared Distances to the Center of Gravity of Each Class: Now, the normalization is defined as:

$$\phi(\mathbf{F}) \triangleq \min_{\mathbf{c}_1} \sum_{i=1}^{N_1} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}_1\|_2^2 + \min_{\mathbf{c}_2} \sum_{i=N_1+1}^{N} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}_2\|_2^2. \tag{30}$$

Similar to the derivation from (27) to (28), it is not hard to rewrite $\phi(\mathbf{F})$ as $\phi(\mathbf{F}) = \mathrm{Tr}(\mathbf{F}^\top \mathbf{F}\mathbf{A}_0)$ where

$$\mathbf{A}_0 \triangleq \sum_{i=1}^{N_1} (\mathbf{x}_i - \bar{\mathbf{x}}_{+1})(\mathbf{x}_i - \bar{\mathbf{x}}_{+1})^\top + \sum_{i=N_1+1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}_{-1})(\mathbf{x}_i - \bar{\mathbf{x}}_{-1})^\top. \tag{31}$$

Remark 5: Note that the normalization (30) is a specific case of (6) with $\boldsymbol{\Theta} = \{0\}$ and the above $\mathbf{A}_0$ defined by (31). For the problem of interest (12) with this newly defined normalization (30), Prop. 2 and Prop. 3 apply for the cases of general $\mathbf{F}$ and diagonal $\mathbf{F}$ respectively.

3) Maximum of Squared Distances to the Center of Gravity of Each Class: The normalization now is defined as:

$$\phi(\mathbf{F}) \triangleq \max \left\{ \min_{\mathbf{c}_1} \sum_{i=1}^{N_1} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}_1\|_2^2, \min_{\mathbf{c}_2} \sum_{i=N_1+1}^{N} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}_2\|_2^2 \right\}. \tag{32}$$

Similar to the derivation from (27) to (28), $\phi(\mathbf{F})$ can be rewritten as $\phi(\mathbf{F}) = \max\{\mathrm{Tr}(\mathbf{F}^\top \mathbf{F}\mathbf{A}_{+1}), \mathrm{Tr}(\mathbf{F}^\top \mathbf{F}\mathbf{A}_{-1})\}$ where

$$\mathbf{A}_{+1} \triangleq \sum_{i=1}^{N_1} (\mathbf{x}_i - \bar{\mathbf{x}}_{+1})(\mathbf{x}_i - \bar{\mathbf{x}}_{+1})^\top, \tag{33}$$

$$\mathbf{A}_{-1} \triangleq \sum_{i=N_1+1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}_{-1})(\mathbf{x}_i - \bar{\mathbf{x}}_{-1})^\top. \tag{34}$$

Remark 6: Note that the normalization (32) is a specific case of (6) with $\boldsymbol{\Theta} = \{+1, -1\}$ and the above $\mathbf{A}_{+1}$ and $\mathbf{A}_{-1}$ defined by (33) and (34). For the problem of interest (12) with this newly defined normalization (32), Prop. 2 and Prop. 3 apply for the cases of general $\mathbf{F}$ and diagonal $\mathbf{F}$ respectively.

C. Normalizations Based on Squared Radius of the Smallest Sphere

1) Squared Radius of the Smallest Sphere Containing All the Training Samples: The radius of the smallest sphere that contains all the transformed training samples is defined as[3]:

$$R_0 \triangleq \min_{\mathbf{c}} \max_{i=1,\ldots,N} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}\|_2. \tag{35}$$

Now, we can define the normalization as:

$$\phi(\mathbf{F}) \triangleq R_0^2. \tag{36}$$

Under this normalization, the normalized margin becomes $\frac{M}{R_0}$. Interestingly, the number of misclassification errors of the leave-one-out error is upper bounded by $\frac{1}{N}\frac{R_0^2}{M^2}$ [3], which provides a nice interpretation for the problem of interest (12).

By finding the dual problem of (36) and after some mathematical manipulations, $\phi(\mathbf{F})$ can be expressed as:

$$\begin{aligned}\phi(\mathbf{F}) &= \max_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}_0} \left\{ \sum_{i=1}^{N} \lambda_{0i} \mathbf{x}_i^\top \mathbf{F}^\top \mathbf{F}\mathbf{x}_i - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{0i}\lambda_{0j} \mathbf{x}_i^\top \mathbf{F}^\top \mathbf{F}\mathbf{x}_j \right\} \\ &= \max_{\boldsymbol{\lambda}_0 \in \boldsymbol{\Lambda}_0} \mathrm{Tr}\left(\mathbf{F}^\top \mathbf{F}\mathbf{A}_{\boldsymbol{\lambda}_0}\right)\end{aligned} \tag{37}$$

where

$$\mathbf{A}_{\boldsymbol{\lambda}_0} \triangleq \sum_{i=1}^{N} \lambda_{0i}\mathbf{x}_i\mathbf{x}_i^\top - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{0i}\lambda_{0j}\mathbf{x}_i\mathbf{x}_j^\top \tag{38}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_{0i}\lambda_{0j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{39}$$

$$\boldsymbol{\Lambda}_0 \triangleq \left\{ \boldsymbol{\lambda}_0 \in \mathbb{R}^N | \boldsymbol{\lambda}_0 \geq \mathbf{0}, \mathbf{1}^\top \boldsymbol{\lambda}_0 = 1 \right\}. \tag{40}$$

[3]If we define R as $R \triangleq \min_{\mathbf{c}} \max_{i=1,\ldots,N} \|\mathbf{F}(\mathbf{x}_i - \mathbf{c})\|_2$, we will get the same result.

Remark 7: Note that the normalization (36) is a specific case of (6) with $\mathbf{A}_{\boldsymbol{\theta}} = \mathbf{A}_{\boldsymbol{\lambda}_0}$ and $\boldsymbol{\Theta} = \boldsymbol{\Lambda}_0$ defined by (38) and (40). For the problem of interest (12) with this newly defined normalization (36), Prop. 2 and Prop. 3 apply for the cases of general $\mathbf{F}$ and diagonal $\mathbf{F}$ respectively.

2) Summation of Squared Radius of the Smallest Sphere Containing the Training Samples Within Each Class: Define the normalization as the summation of the squared radius of the smallest spheres that contain training samples within each class, that is

$$\phi(\mathbf{F}) \triangleq R_{+1}^2 + R_{-1}^2 \tag{41}$$

where

$$R_{+1} \triangleq \min_{\mathbf{c}} \max_{i=1,\ldots,N_1} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}\|_2, \tag{42}$$

$$R_{-1} \triangleq \min_{\mathbf{c}} \max_{i=N_1+1,\ldots,N} \|\mathbf{F}\mathbf{x}_i - \mathbf{c}\|_2. \tag{43}$$

Similar to (37), easily we can rewrite $\phi(\mathbf{F})$ as:

$$\begin{aligned}\phi(\mathbf{F}) &= \max_{\boldsymbol{\lambda}_{+1}\in\boldsymbol{\Lambda}_{+1}} \mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\lambda}_{+1}}\right) + \max_{\boldsymbol{\lambda}_{-1}\in\boldsymbol{\Lambda}_{-1}} \mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\lambda}_{-1}}\right)\\ &= \max_{(\boldsymbol{\lambda}_{+1},\boldsymbol{\lambda}_{-1})\in\boldsymbol{\Lambda}_{+1}\times\boldsymbol{\Lambda}_{-1}} \mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\left(\mathbf{A}_{\boldsymbol{\lambda}_{+1}} + \mathbf{A}_{\boldsymbol{\lambda}_{-1}}\right)\right)\end{aligned}$$

where

$$\mathbf{A}_{\boldsymbol{\lambda}_{+1}} \triangleq \frac{1}{2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_1}\lambda_{+1i}\lambda_{+1j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top, \tag{44}$$

$$\begin{aligned}\mathbf{A}_{\boldsymbol{\lambda}_{-1}} \triangleq \frac{1}{2}\sum_{i=N_1+1}^{N}\sum_{j=N_1+1}^{N}&\left[\lambda_{-1(i-N_1)}\lambda_{-1(j-N_1)}\right.\\ &\left.\times(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top\right],\end{aligned} \tag{45}$$

$$\boldsymbol{\Lambda}_{+1} \triangleq \left\{\boldsymbol{\lambda}_{+1}\in\mathbb{R}^{N_1} | \boldsymbol{\lambda}_{+1} \geq \mathbf{0},\ \mathbf{1}^\top\boldsymbol{\lambda}_{+1} = 1\right\}, \tag{46}$$

$$\boldsymbol{\Lambda}_{-1} \triangleq \left\{\boldsymbol{\lambda}_{-1}\in\mathbb{R}^{N-N_1} | \boldsymbol{\lambda}_{-1} \geq \mathbf{0}, \mathbf{1}^\top\boldsymbol{\lambda}_{-1} = 1\right\}. \tag{47}$$

Remark 8: Note that the normalization (41) is a specific case of (6) with $\boldsymbol{\Theta} = \boldsymbol{\Lambda}_{+1}\times\boldsymbol{\Lambda}_{-1}$ and $\mathbf{A}_{\boldsymbol{\theta}} = \mathbf{A}_{\boldsymbol{\lambda}_{+1}} + \mathbf{A}_{\boldsymbol{\lambda}_{-1}}$ where the sets and matrices are defined by (44)–(47). For the problem of interest (12) with this newly defined normalization (41), Prop. 2 and Prop. 3 apply for the cases of general $\mathbf{F}$ and diagonal $\mathbf{F}$ respectively.

3) Maximum of Squared Radius of the Smallest Sphere Containing the Training Samples Within Each Class: Now, instead of considering summation in (41), we move to the maximum and define $\phi(\mathbf{F})$ as:

$$\phi(\mathbf{F}) \triangleq \max\left\{R_{+1}^2, R_{-1}^2\right\} \tag{48}$$

and similar to (37), it can be rewritten as:

$$\begin{aligned}\phi(\mathbf{F}) = \max\Bigg\{&\max_{\boldsymbol{\lambda}_{+1}\in\boldsymbol{\Lambda}_{+1}} \mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\lambda}_{+1}}\right),\\ &\max_{\boldsymbol{\lambda}_{-1}\in\boldsymbol{\Lambda}_{-1}} \mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\lambda}_{-1}}\right)\Bigg\}\end{aligned} \tag{49}$$

where $\mathbf{A}_{\boldsymbol{\lambda}_{+1}}$, $\mathbf{A}_{\boldsymbol{\lambda}_{-1}}$, $\boldsymbol{\Lambda}_{+1}$ and $\boldsymbol{\Lambda}_{-1}$ are defined by (44)–(47).

Note that the normalization (49) is a little bit more complicated than (6). However, for the problem (12) with the above normalization (49), similar to Prop. 2 and Prop. 3 in Section IV, we can have the following results.

Corollary 1: Let $\mathbf{F}$ be a general matrix, the problem (12) with the above definition of $\phi(\mathbf{F})$ in (49) is bounded by:

$$\begin{aligned}\underset{\mathbf{v},\beta_0,\boldsymbol{\xi}}{\text{minimize}}\quad & \frac{1}{2}\psi(\mathbf{v}) + C\mathbf{1}^\top\boldsymbol{\xi}\\ \text{subject to}\quad & y_i(\mathbf{v}^\top\mathbf{x}_i + \beta_0) \geq 1 - \xi_i, \quad \forall i\\ & \boldsymbol{\xi} \geq \mathbf{0},\end{aligned} \tag{50}$$

where

$$\psi(\mathbf{v}) \triangleq \max\left\{\max_{\boldsymbol{\lambda}_{+1}\in\boldsymbol{\Lambda}_{+1}} \frac{1}{2}\left\|\mathbf{A}_{\boldsymbol{\lambda}_{+1}}^{1/2}\mathbf{v}\right\|_2^2, \max_{\boldsymbol{\lambda}_{-1}\in\boldsymbol{\Lambda}_{-1}} \frac{1}{2}\left\|\mathbf{A}_{\boldsymbol{\lambda}_{-1}}^{1/2}\mathbf{v}\right\|_2^2\right\}$$

and $\mathbf{A}_{\boldsymbol{\lambda}_{+1}}$, $\mathbf{A}_{\boldsymbol{\lambda}_{-1}}$, $\boldsymbol{\Lambda}_{+1}$ and $\boldsymbol{\Lambda}_{-1}$ are defined by (44)–(47). In addition, if there exists either some $\boldsymbol{\lambda}_{+1} \in \boldsymbol{\Lambda}_{+1}$ with $\mathbf{A}_{\boldsymbol{\lambda}_{+1}}$ being full rank or some $\boldsymbol{\lambda}_{-1} \in \boldsymbol{\Lambda}_{-1}$ with $\mathbf{A}_{\boldsymbol{\lambda}_{-1}}$ being full rank, the lower bound is tight.

Proof: The proof is only a slight extension of that of Prop. 2 and thus omitted. ∎

Corollary 2: Let $\mathbf{F}$ be diagonal, the problem (12) with definition of $\phi(\mathbf{F})$ in (49) is bounded by:

$$\begin{aligned}\underset{\mathbf{v},\beta_0,\boldsymbol{\xi},t}{\text{minimize}}\quad & \frac{1}{2}\tilde{\psi}(\mathbf{v}) + C\mathbf{1}^\top\boldsymbol{\xi}\\ \text{subject to}\quad & y_i(\mathbf{v}^\top\mathbf{x}_i + \beta_0) \geq 1 - \xi_i, \quad \forall i\\ & \boldsymbol{\xi} \geq \mathbf{0},\end{aligned} \tag{51}$$

where

$$\begin{aligned}\tilde{\psi}(\mathbf{v}) \triangleq \max\Bigg\{&\max_{\boldsymbol{\lambda}_{+1}\in\boldsymbol{\Lambda}_{+1}} \frac{1}{2}\left\|\mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\lambda}_{+1}}\right)\mathbf{v}\right\|_1^2,\\ &\max_{\boldsymbol{\lambda}_{-1}\in\boldsymbol{\Lambda}_{-1}} \frac{1}{2}\left\|\mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\lambda}_{-1}}\right)\mathbf{v}\right\|_1^2\Bigg\}\end{aligned}$$

and $\mathbf{A}_{\boldsymbol{\lambda}_{+1}}$, $\mathbf{A}_{\boldsymbol{\lambda}_{-1}}$, $\boldsymbol{\Lambda}_{+1}$ and $\boldsymbol{\Lambda}_{-1}$ are defined by (44)–(47). In addition, if there exists either some $\boldsymbol{\lambda}_{+1} \in \boldsymbol{\Lambda}_{+1}$ with $\mathrm{Diag}(\mathbf{A}_{\boldsymbol{\lambda}_{+1}})$ being full rank or some $\boldsymbol{\lambda}_{-1} \in \boldsymbol{\Lambda}_{-1}$ with $\mathrm{Diag}(\mathbf{A}_{\boldsymbol{\lambda}_{-1}})$ being full rank, the lower bound is tight.

Proof: The proof is only a slight extension of that of Prop. 3 and thus omitted. ∎

VI. Proposed Unified Framework

Based on the general linear SVM formulation (12), the solving approach, and the explorations in Sections IV, V, finally we are able to reach our unified framework of SVM, i.e., problem (UF), as stated before in Table I. The proposed unified framework is controlled by the vector norm square $\|\cdot\|_{1\text{ or }2}^2$, and $\mathbf{A}_{\boldsymbol{\theta}}$, which determines the weights of the vector norm, and $\boldsymbol{\Theta}$ which takes pointwise maximum over different weighted vector norms.

In Table I, each row shows one type of (UF), and each combination of some type of (UF) represents one normalization. For example, Example 1 with $\mathbf{F}$ being diagonal mentioned in Section IV is the combination of type 1 with $\boldsymbol{\Theta} = \{0\}$, ℓ_1-norm, and $\mathbf{A}_0$ given by (9), and the different normalizations having been explored in Section V can also easily be found as different combinations of some types.

Mathematically speaking, we can simply characterize (or propose) the different existing (or new) SVM formulations based on this unified framework by trying different combinations of norm, $\boldsymbol{\Theta}$ and $\mathbf{A}_{\boldsymbol{\theta}}$. The origin of having such different methods can be nicely interpreted as taking different normalizations $\phi(\mathbf{F})$ in the general linear SVM (12).

In the following part of this section, we will explore a lot of existing SVMs as the specific cases under the proposed unified framework, investigate the insights under different normalizations, and furthermore propose some more SVMs based on the unified framework. This is also the benefit of having such a general unified framework.

A. Existing SVMs as Special Cases

1) Standard Soft-Margin SVM: Consider the normalization

$$\phi(\mathbf{F}) \triangleq \mathrm{Tr}(\mathbf{F}^\top \mathbf{F}) \tag{52}$$

where

$$\mathbf{A}_0 \triangleq \mathbf{I}. \tag{53}$$

According to Prop. 2, we can easily recover the standard soft-margin SVM (2), which corresponds to a combination with ℓ_2-norm and $\mathbf{A}_0$ given by (53) of type 1 in Table I. Note that $\phi(\mathbf{F}) = \mathrm{Tr}(\mathbf{F}^\top \mathbf{F})$ does not take any information of the training data into account.

2) ℓ_1-Norm SVM: Still, we insist on using the normalization (52), however we set $m = d$ and restrict $\mathbf{F}$ to be diagonal. From Prop. 3, we recover the ℓ_1-SVM (3) as a combination with ℓ_1-norm and $\mathbf{A}_0$ given by (53) of type 1 in Table I.

3) SVM With Standardization: One technique often used in the literature is to standardize the raw data first and then apply SVM to the standardized data. This can be formulated as:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i\left(\boldsymbol{\beta}^\top \mathbf{S}\left(\mathbf{x}_i - \mathbf{c}\right) + \beta_0\right) \geq 1 - \xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}, \end{array} \tag{54}$$

where $\mathbf{c}$ denotes the mean of the sequence $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and $\mathbf{S}$ is diagonal matrix with $1/\mathbf{S}_{ii}$ denoting the sample standard deviation[4] of the i-th feature. Obviously, problem (54) is equivalent to:

$$\begin{array}{ll} \underset{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i(\boldsymbol{\beta}^\top \mathbf{S}\mathbf{x}_i + \beta_0) \geq 1 - \xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}, \end{array} \tag{55}$$

which can be further reformulated as:

$$\begin{array}{ll} \underset{\mathbf{v},\beta_0,\boldsymbol{\xi}}{\text{minimize}} & \frac{1}{2}\left\|\mathbf{A}^{1/2}\mathbf{v}\right\|_2^2 + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} & y_i(\mathbf{v}^\top \mathbf{x}_i + \beta_0) \geq 1 - \xi_i, \quad \forall i \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{array} \tag{56}$$

where $\mathbf{A}$ is a diagonal matrix with $\mathbf{A}_{ii} = 1/\mathbf{S}_{ii}^2$, i.e., the sample variance of the i-th attribute.

Another interesting observation is that, if we define the normalization:

$$\phi(\mathbf{F}) \triangleq \mathrm{Tr}(\mathbf{F}^\top \mathbf{F}\mathbf{A}_0) \tag{57}$$

where

$$\mathbf{A}_0 \triangleq \mathrm{Diag}\left(\mathsf{Cov}\left([\mathbf{x}_1 \cdots \mathbf{x}_N]\right)\right), \tag{58}$$

then from Prop. 2, we can recover problem (56), or equivalently (54), from the general problem (12) as a combination with ℓ_2-norm and $\mathbf{A}_0$ given by (58) of type 1 in Table I since $\mathbf{A}_{0ii} = 1/\mathbf{S}_{ii}^2$ indeed.

Similarly, still using (57), the ℓ_1-SVM with standardization can be recovered as a combination with ℓ_1-norm and $\mathbf{A}_0$ given by (58) of type 1 in Table I according to Prop. 3.

Unlike $\mathrm{Tr}(\mathbf{F}^\top \mathbf{F})$ in (52), the normalization (57) does make use of the training data.

Remark 9: Note that

$$\mathsf{Cov}\left([\mathbf{x}_1 \cdots \mathbf{x}_N]\right) = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}}_0)(\mathbf{x}_i - \bar{\mathbf{x}}_0)^\top \tag{59}$$

is equal to (29) up to a scalar $\frac{1}{N-1}$. This means that (58) actually only considers the diagonal elements of (29) where the normalization is based on the distances to the center of gravity. Under the proposed framework, easily we discover the connection between the two SVMs, i.e., (54) and (12) with $\phi(\mathbf{F})$ defined by (27), via the link between their normalizations, that is, (58) is the diagonal part of (29).

Remark 10: Apart from the above standardization method (usually called z-score method also), there are many other standardization methods, say scaling all the attribute values among [0,1]. A more detailed description on different standardizations can be found in [13], [14]. Similar to the above z-score method, all the other standardization methods can be cast into our proposed framework with some specific matrix $\mathbf{A}_{\boldsymbol{\theta}}$. Since it is quite easy and straightforward, we omit it here.

B. More Insights Based on the Unified Framework

Insights From Energy Point of View: Observe the normalization definition $\phi(\mathbf{F}) \triangleq \mathrm{Tr}(\mathbf{F}^\top \mathbf{F})$ in (52) for standard soft-margin SVM (2) and ℓ_1-SVM (3), it can also be treated as total energy of the transformation $\mathbf{F}$. And the inequality constraint $\phi(\mathbf{F}) \leq 1$ in formulation (57) can be interpreted as an upper bound on the energy, which is quite similar to the power constraint in the precoder design problems in communication systems [21], [22]. Thus, the methods standard soft-margin SVM (2) and ℓ_1-SVM (3) evaluate the energy of the transformation of different features uniformly, i.e., they do not evaluate the energy in the transformation of one feature more important than that of another feature. However, the energy measure of the SVM with standardization in (57) puts more weights on energy in the transformation of the features with larger variances. One advantage of (57) compared with $\mathrm{Tr}(\mathbf{F}^\top \mathbf{F})$ is that the weights can capture some data structure information. Similarly, the methods we have reviewed and derived in Section IV and V can be interpreted as different ways, i.e., different $\phi(\mathbf{F})$, to evaluate the energy of the transformation $\mathbf{F}$ based on the training data.

Insights From Penalty Point of View: If we look at the term $\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\|\mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}\|_{1 \text{ or } 2}^2\}$ in the unified framework, different combinations of norms and $\mathbf{A}_{\boldsymbol{\theta}}$ define different types of penalty on the features. Thus both the soft-margin SVM (2) and ℓ_1-SVM (3) do not consider the information in the training data and penalize different features equally, but they still differ from each other by adopting different vector norms. However, the SVM

[4]Note that we implicitly assume that the sample standard deviation of each feature greater than zero. Otherwise, the feature does not affect the classification and thus can be removed.

with standardization uses the $\mathrm{Diag}(\mathrm{Cov}([\mathbf{x}_1 \cdots \mathbf{x}_N]))$ in (57) as penalty weights. Similarly, the methods we have covered in Section IV and V also provide different types of penalty, in which the vector norm depends on the transform mapping $\mathbf{F}$, and the weights matrix $\mathbf{A}_{\boldsymbol{\theta}}$ depends on how we look at the training data. All the above different types of penalty result from using different normalizations $\phi(\mathbf{F})$.

Since it is usually easier to gain geometric insight from the ℓ_2-norm, we investigate two examples with ℓ_2-norm. For the case that $\boldsymbol{\Theta}$ and $\mathbf{A}_0$ are given by (8) and (9), the penalty is

$$\begin{aligned}&\left\|\mathbf{A}_0^{1/2}\mathbf{v}\right\|_2^2\\&=\sum_{i,j=1}^{N}\frac{1+y_iy_j}{2}\left(\mathbf{v}^\top\mathbf{x}_i-\mathbf{v}^\top\mathbf{x}_j\right)\left(\mathbf{v}^\top\mathbf{x}_i-\mathbf{v}^\top\mathbf{x}_j\right)\\&=\sum_{i,j=1}^{N_1}\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right)^2+\sum_{i,j=N_1+1}^{N}\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right)^2\end{aligned}\tag{60}$$

where $f(\mathbf{x})=\mathbf{v}^\top\mathbf{x}+\beta_0$ is defined in (1) and $f(\mathbf{x})=0$ represents the separating hyperplane. Since $f(\mathbf{x}_i)$ represents the signed distance of point $\mathbf{x}_i$ to the separating hyperplane up to a common positive scalar,[5] thus (60) penalizes the summation of squared differences among the distances of the samples to the separating hyperplane within each class.

Now, if we consider $\boldsymbol{\Theta}=\{+1,-1\}$ and $\mathbf{A}_{+1}$ and $\mathbf{A}_{-1}$ are given by (33) and (34) separately, then the penalty is

$$\begin{aligned}&\max\left\{\left\|\mathbf{A}_{+1}^{1/2}\mathbf{v}\right\|_2^2,\left\|\mathbf{A}_{-1}^{1/2}\mathbf{v}\right\|_2^2\right\}=\max\left\{\mathbf{v}^\top\mathbf{A}_{+1}\mathbf{v},\mathbf{v}^\top\mathbf{A}_{-1}\mathbf{v}\right\}\\&=\max\left\{\sum_{i=1}^{N_1}\left(f(\mathbf{x}_i)-f\left(\bar{\mathbf{x}}_{+1}\right)\right)^2,\right.\\&\qquad\left.\sum_{i=N_1+1}^{N}\left(f(\mathbf{x}_i)-f\left(\bar{\mathbf{x}}_{-1}\right)\right)^2\right\}\end{aligned}\tag{61}$$

where $f(\mathbf{x})=\mathbf{v}^\top\mathbf{x}+\beta_0$ is defined in (1) and $f(\mathbf{x})=0$ represents the separating hyperplane, and $\bar{\mathbf{x}}_{+1}$ and $\bar{\mathbf{x}}_{-1}$ are the means of samples of the class $+1$ and -1 as defined in (21). Thus, (61) penalizes the maximum of the summation of the squared differences between the signed distance of some sample to the separating hyperplane and the signed distance of the mean of the same class to the separating hyperplane.

The underlying ideas of the above two penalty examples 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, i.e., (60) and (61), to measure the compactness of the data.

Intuitively, the above two examples make sense when the training data is somehow flat for both classes along some direction, 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, or equivalently penalizing $\mathbf{v}^\top\mathbf{v}$ only, may give us the separating hyperplane that has wrong direction and could be really misleading for the testing. However, once we have taken the data information into the consideration of the formulation, i.e., penalizing (60) or (61) instead, we can get better separating hyperplane that is closer to the true one. We will use numerical examples to illustrate the insights clearly in Section VII and the numerical experiments show that the proposed methods perform really well for a large range of data sets with different feature label distributions not limited to the flat ones.

Now it is quite straightforward to gain the insights from the penalty point of view for the other normalizations explored in this paper. Since the procedures are really similar to that for the above two examples and thus are omitted.

C. Combinations of Normalizations

Similar to the maximum or summation normalizations we have explored in Section V, easily we can have more various normalizations by combining different normalizations.

1) Maximum of Normalizations: First of all, we can generally have more normalizations as the maximum of two normalizations:

$$\phi(\mathbf{F})\triangleq\max_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_1,\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_2}\left\{\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}_1}\right),\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}_2}\right)\right\}$$

where $\mathbf{A}_{\boldsymbol{\theta}_1}$ and $\mathbf{A}_{\boldsymbol{\theta}_2}$ can be any weight matrices, and this can be easily extended to more than two normalizations.

2) Summation of Normalizations: Another thing is that we can also easily obtain a new normalization term simply by adding two existing normalizations:

$$\begin{aligned}\phi(\mathbf{F})&\triangleq\max_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_1,\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_2}\left\{\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}_1}\right)+\nu\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}_2}\right)\right\}\\&=\max_{\boldsymbol{\theta}_1\in\boldsymbol{\Theta}_1,\boldsymbol{\theta}_2\in\boldsymbol{\Theta}_2}\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\left(\mathbf{A}_{\boldsymbol{\theta}_1}+\nu\mathbf{A}_{\boldsymbol{\theta}_2}\right)\right),\end{aligned}\tag{62}$$

where $\mathbf{A}_{\boldsymbol{\theta}_1}$ and $\mathbf{A}_{\boldsymbol{\theta}_2}$ can be any weight matrices, and $\nu\geq 0$ is the trade-off parameter. Note that if $\nu=0$ we only considers the first normalization and if $\nu=\infty$ we only considers the second normalization. Also, obviously this can be extended to the summation of more than two normalizations.

An interesting case is to combine $\mathrm{Tr}(\mathbf{F}^\top\mathbf{F})$ with some other normalization:

$$\phi(\mathbf{F})\triangleq\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\mathrm{Tr}\left(\mathbf{F}^\top\mathbf{F}\mathbf{A}_{\boldsymbol{\theta}}\right)+\nu\mathrm{Tr}(\mathbf{F}^\top\mathbf{F}).\tag{63}$$

Here, $\mathbf{A}_{\boldsymbol{\theta}}$ can be any weight matrix. For example, if $\boldsymbol{\Theta}$ and $\mathbf{A}_{\boldsymbol{\theta}}$ are given by (8) and (9) and ℓ_2-norm is used, similar to (60), normalization (63) gives the penalty:

$$\left\|(\mathbf{A}_0+\nu\mathbf{I})^{1/2}\mathbf{v}\right\|_2^2=\sum_{i,j=1}^{N}\frac{1+y_iy_j}{2}\left(f(\mathbf{x}_i)-f(\mathbf{x}_j)\right)^2+\nu\mathbf{v}^\top\mathbf{v}\tag{64}$$

where the quantity $\sum_{i,j=1}^{N}\frac{1+y_iy_j}{2}(f(\mathbf{x}_i)-f(\mathbf{x}_j))^2$ can be interpreted as the measurement how compact the samples within each class are with respect to the separating hyperplane, and $\frac{1}{\sqrt{\mathbf{v}^\top\mathbf{v}}}$ stands for the soft-margin. Thus, penalizing (64) means finding the trade-off between concrete compaction of the samples (i.e., small $\sum_{i,j=1}^{N}\frac{1+y_iy_j}{2}(f(\mathbf{x}_i)-f(\mathbf{x}_j))^2$) and large soft-margin (i.e., small $\mathbf{v}^\top\mathbf{v}$).

The above new normalizations show that the proposed unified framework can benefit us with more optional SVMs. Moreover, we need to point out that the above newly proposed combinations of normalizations are only some examples. It is actually quite simple and straightforward for us to have much more

[5] Actually, $(\mathbf{v}^\top\mathbf{x}_i+\beta_0)/\|\mathbf{v}\|_2$ is exactly the signed distance from point $\mathbf{x}_i$ to the hyperplane $\mathbf{v}^\top\mathbf{x}+\beta_0=0$, see [6, Eq. (4.40)].

TABLE II
THE SIMULATED SVMS. HERE $[\mathbf{X}]_{F=d} \triangleq \sqrt{d}\frac{\mathbf{X}}{\|\mathbf{X}\|_F}$ SO THAT $\|[\mathbf{X}]_{F=d}\|_F = \sqrt{d}$

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

Method	$\boldsymbol{\Theta}$	$\|\cdot\|_{1\text{ or }2}^2$	$\mathbf{A}_{\boldsymbol{\theta}}$	Note
I	$\{0\}$	$\|\cdot\|_1^2$	$\mathbf{A}_0 = \left[\text{Diag}\left(\sum_{i,j=1}^N \frac{1+y_iy_j}{2}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top\right)\right]_{F=d}$	Existing
II	$\{0\}$	$\|\cdot\|_2^2$	$\mathbf{A}_0 = \mathbf{I}$	
III	$\{0\}$	$\|\cdot\|_1^2$	$\mathbf{A}_0 = \mathbf{I}$	
IV	$\{0\}$	$\|\cdot\|_2^2$	$\mathbf{A}_0 = \left[\sum_{i,j=1}^N \frac{1+y_iy_j}{2}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top\right]_{F=d} + \nu\mathbf{I}$	Proposed
V	$\{+1,-1\}$	$\|\cdot\|_2^2$	$\mathbf{A}_{+1} = \left[\sum_{i,j=1}^{N_1}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top\right]_{F=d} + \nu_1\mathbf{I}$ $\mathbf{A}_{-1} = \left[\sum_{i,j=N_1+1}^{N}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^\top\right]_{F=d} + \nu_2\mathbf{I}$	
VI	$\{0\}$	$\|\cdot\|_2^2$	$\mathbf{A}_0 = \left[\sum_{i=1}^N(\mathbf{x}_i-\bar{\mathbf{x}}_0)(\mathbf{x}_i-\bar{\mathbf{x}}_0)^\top\right]_{F=d} + \nu\mathbf{I}$	
VII	$\{+1,-1\}$	$\|\cdot\|_2^2$	$\mathbf{A}_{+1} = \left[\sum_{i=1}^{N_1}(\mathbf{x}_i-\bar{\mathbf{x}}_{+1})(\mathbf{x}_i-\bar{\mathbf{x}}_{+1})^\top\right]_{F=d} + \nu_1\mathbf{I}$ $\mathbf{A}_{-1} = \left[\sum_{i=N_1+1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}}_{-1})(\mathbf{x}_i-\bar{\mathbf{x}}_{-1})^\top\right]_{F=d} + \nu_2\mathbf{I}$	

SVMs based on our proposed unified framework, and this shows the great potential wide applications of the proposed unified framework on different kinds of data sets.

D. Discussion on Extensions to Kernel SVMs

The extension of the idea of "normalized margin" and the following unified framework from linear SVM to kernel versions in general is not easy. Given the kernel, one way is to follow the heuristic in [19, Section 5] to construct a nonlinear mapping which induces the same sample training and testing kernels, and then the linear SVM with "normalized margin" (or equivalently, the proposed unified framework) in this paper can be applied as usual to the mapped training and testing data.

VII. NUMERICAL EXPERIMENTS

A. Simulated Methods

Now we have all the reviewed and proposed SVMs summarized as the combinations in Table I. However, there are too many combinations and some of them are quite similar. To keep the numerical experiments simple while illustrating the benefits of the proposed framework clearly, we evaluate only some combinations via numerical examples based on both synthetic data and real-world data from different sources. Basically, we will mainly study how the proposed SVMs, like the ones with penalties (60), (61), (62), and (64), etc., perform compared to the existing ones, e.g., the traditional ℓ_1-norm, ℓ_2-norm SVMs and the one proposed in [19], especially when the number of training samples is not large enough. Table II summarizes the seven SVMs we will simulate named methods I–VII respectively in this section. All the optimization problems are solved via the commercial solver MOSEK [26] in MATLAB.

Before we have shown that the proposed unified framework is able to take the data information into the problem formulation, e.g., see Remark 1. Here we also want to compare the methods in Table II with the classical linear discriminate analysis (LDA) method, which also uses data information.

For the binary classification problem, the LDA finds [6]

$$\mathbf{v}_{\text{LDA}} \triangleq \bar{\boldsymbol{\Sigma}}^{-1}(\bar{\mathbf{x}}_{+1}-\bar{\mathbf{x}}_{-1}) \tag{65}$$

$$\beta_{0\text{LDA}} \triangleq \frac{1}{2}\bar{\mathbf{x}}_{-1}^\top\bar{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{x}}_{-1} - \frac{1}{2}\bar{\mathbf{x}}_{+1}^\top\bar{\boldsymbol{\Sigma}}^{-1}\bar{\mathbf{x}}_{+1} + \log\left(\frac{N_1}{N}\right) - \log\left(1-\frac{N_1}{N}\right) \tag{66}$$

where

$$\bar{\boldsymbol{\Sigma}} \triangleq \frac{1}{N-2}\left[\sum_{i=1}^{N_1}(\mathbf{x}_i-\bar{\mathbf{x}}_{+1})(\mathbf{x}_i-\bar{\mathbf{x}}_{+1})^\top + \sum_{i=N_1+1}^{N}(\mathbf{x}_i-\bar{\mathbf{x}}_{-1})(\mathbf{x}_i-\bar{\mathbf{x}}_{-1})^\top\right] \tag{67}$$

and the classification prediction for a new outcome sample $\mathbf{x}$ is $\hat{y} = \text{sign}(\mathbf{v}_{\text{LDA}}^\top\mathbf{x} + \beta_{0\text{LDA}})$. The interpretation of the LDA is that $\mathbf{v}_{\text{LDA}}$ is the direction along which the ratio of the between-class variance to the within-class variance of the projections of $\mathbf{x}_1,\ldots,\mathbf{x}_N$ is maximized. That is, $\mathbf{v}_{\text{LDA}}$ maximizes $\frac{\mathbf{v}^\top\mathbf{S}_\text{b}\mathbf{v}}{\mathbf{v}^\top\mathbf{S}_\text{w}\mathbf{v}}$ where $\mathbf{S}_\text{b} \triangleq (\bar{\mathbf{x}}_{+1}-\bar{\mathbf{x}}_{-1})(\bar{\mathbf{x}}_{+1}-\bar{\mathbf{x}}_{-1})^\top$ is called the between-class scatter matrix and $\mathbf{S}_\text{w} \triangleq (N-2)\bar{\boldsymbol{\Sigma}}$ is called the within-class scatter matrix.

B. Experiments on Synthetic Data

To illustrate the insights clearly, we consider visualizable synthetic experiments with only two attributes, that is, $d = 2$. Here we consider two classes with equal probabilities, and their samples are drawn randomly from two Gaussian distributions: samples of class $+1$ are drawn from $\mathcal{N}(\mathbf{1}_d, \boldsymbol{\Sigma}_{+1})$ and samples of

class -1 are drawn from $\mathcal{N}(-\mathbf{1}_d, \mathbf{\Sigma}_{-1})$, where $\mathbf{\Sigma}_{+1}, \mathbf{\Sigma}_{-1} \in \mathbb{R}^{d\times d}$ are the covariance matrices.

1) Same Covariance Matrices: First, we consider the case of equal covariance matrices:

$$\mathbf{\Sigma}_{+1} = \mathbf{\Sigma}_{-1} = 4 \times \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}.$$

Under the Gaussian distribution condition and the two classes have the same covariance matrices, the theoretical LDA provides the optimal decision boundary of those two classes [6]. For this simulated case, the optimal decision boundary is $f(\mathbf{x}) = \mathbf{1}_d^\top \mathbf{x} = x_1 + x_2 = 0$, and the class label prediction of $\mathbf{x}$ simply is $\hat{y} = \text{sign}(f(\mathbf{x}))$. Later in Fig. 2(a), we can see the two classes (see the green ellipses as contours) and the optimal boundary (see the red solid line).

Now we set up the simulations. For each realization, the two classes $+1$ and -1 have equal probabilities. Once the class is determined, we can generate the data from the aforementioned corresponding Gaussian distributions. We randomly generate 1000 samples as the synthetic data set. Since we focus on the scenarios that the training sample size is not large, to avoid the case that there is no training sample for one class, we randomly draw training samples from both classes with equal numbers. Say the training sample size is $N = 10$, then we randomly select 5 from both classes $+1$ and -1, and the remaining samples are used for testing. Then, we simply run all the SVMs listed in Table II for the tuning parameter $C \in \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$. To avoid that the performance is biased by one realization, we repeat the above realization for 100 times, compute average test error rate for each tuning parameter C for each SVM, and report the best average test error rate for each method. For methods IV–VII, we set $\nu = \nu_1 = \nu_2 \in \{0.1, 0.5, 1, 1.5\}$ and keep the best result.

Intuitively, for the linearly nonseparable Gaussian distributions, all the methods should work pretty well and perform quite closely to each other when there are enough training samples because all the SVMs have enough training samples and the learnt separating hyperplanes will be somehow close to the true one. However, when there are not enough samples, the proposed SVMs which take the data information into consideration may outperform the existing ones which do not consider data information. Fig. 1 shows the results of the average test error rate versus the number of training samples. Obviously, we can see that the proposed methods outperform the existing ones, especially when the training sample size is relatively small, which coincides with the above intuition and the insights we explored before in Section VI-B as well. Thus, we are more interested in the (relatively) small sample regime.

To understand why the proposed SVMs perform better than the existing ones when the training size is relatively small, we study two specific realizations in Fig. 2. The first realization in Fig. 2(a) shows that the existing method II aiming at maximizing the soft-margin may be misleading as shown as the blue dotted line since only a few support vectors matter and the other samples do not affect the separating hyperplane at all. Fortunately, this drawback can be overcome by the proposed methods, see the separating black dash-dotted line and magenta dashed line in Fig. 2(a) found by the proposed methods IV and VI separately. Even though the theoretical LDA is optimal, however, the empirical LDA does not provide good classifier because of small training set, see the red dotted line in Fig. 2(a). Fig. 2(b) shows

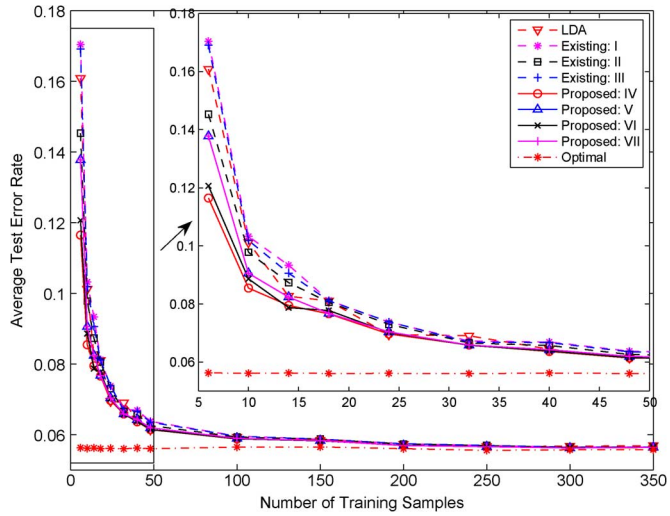

Fig. 1. Synthetic example with same covariance matrices: average test error rate versus number of training samples.

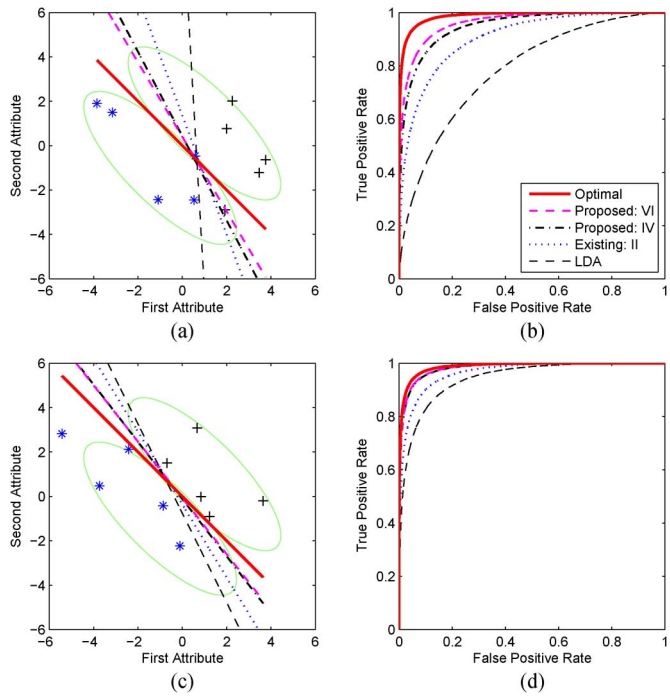

Fig. 2. Two specific synthetic realizations with same covariance matrices: existing SVM II may be misleading when there are not enough training samples (blue dotted line), this drawback can be overcome by the proposed methods (black dash-dotted line and magenta dashed line). (a) Data and classifiers. (b) ROC curves. (c) Data and classifiers. (d) ROC curves.

the Receiver Operating Characteristic (ROC) curves of different methods. We can clearly see that the two proposed methods outperform the standard soft-margin SVM and the LDA. The insight is that 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 w.r.t. the separating line at the same time, just what we have explained before in Section VI-B and VI-C. The second realization in Fig. 2(c) is somehow more ordinary, however, we can still observe that the proposed methods slightly outperform the existing method II and the LDA based on the ROC curves in Fig. 2(d).

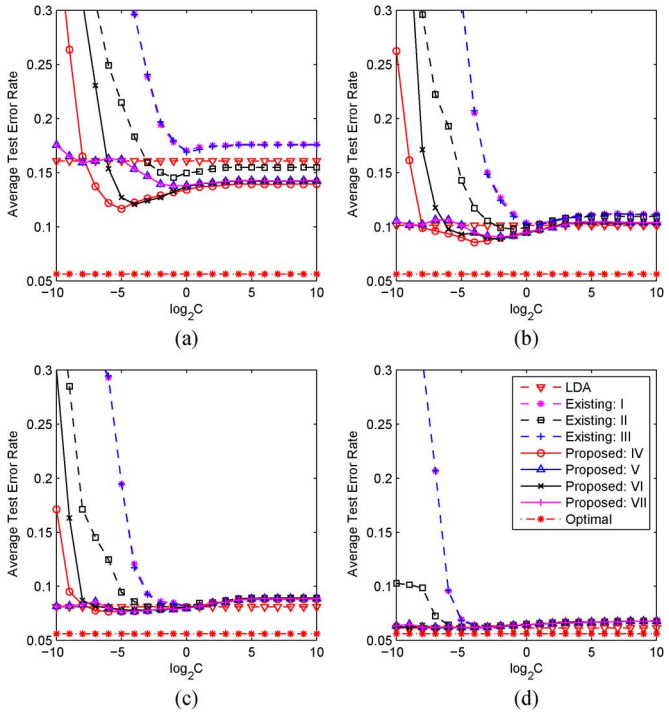

Fig. 3. Synthetic example with same covariance matrices: average test error rate versus tuning parameter C. (a) N $=$ 6. (b) N $=$ 10. (c) N $=$ 18. (d) N $=$ 48.

As to the effect of the tuning parameter C, Fig. 3 shows the average test error rate versus the tuning parameter C for different number of training samples, e.g., $N = 6, 10, 18, 48$. We have several interesting observations. First, the proposed SVMs tend to outperform the existing SVMs over all the C values and the LDA for many C values, especially when N is small (e.g., see Fig. 3(a)). Second, when N is larger, all the methods tend to perform better and more closely. Third, interestingly, one more advantage of the proposed methods we can observe from Fig. 3 is that they are relatively more robust to the tuning parameter C than the existing ones. Fourth, note that the optimal classifier (i.e., the theoretical LDA) and the empirical LDA are independent of C, and when N gets larger, the empirical LDA should be closer to the theoretical LDA and it is verified in Fig. 3 as we look at Figs. 3(a)–(d), the empirical LDA methods performs more and more closely to the optimal method when N grows from 6 to 48.

2) Different Covariance Matrices: Now we move to the case of different covariance matrices:

$$\mathbf{\Sigma}_{-1} = 2 \times \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$

and

$$\mathbf{\Sigma}_{+1} = 1.2 \times \mathbf{R}(-45^\circ) \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix} (\mathbf{R}(-45^\circ))^\top$$

where $\mathbf{R}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ is a rotation matrix. Since $\mathbf{\Sigma}_{+1} \neq \mathbf{\Sigma}_{-1}$, the theoretical LDA is no longer optimal. Actually, by discriminate analysis [6], the optimal decision boundary is quadratic given by $f(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{\Sigma}_{+1}| - \frac{1}{2}(\mathbf{x}-\mathbf{1}_d)^\top \mathbf{\Sigma}_{+1}^{-1}(\mathbf{x}-\mathbf{1}_d) + \log(N_1/N) + \frac{1}{2}\log|\mathbf{\Sigma}_{-1}| + \frac{1}{2}(\mathbf{x}+\mathbf{1}_d)^\top \mathbf{\Sigma}_{-1}^{-1}(\mathbf{x}+\mathbf{1}_d) - \log(1-N_1/N) = 0$, and the class label prediction of $\mathbf{x}$ simply is $\hat{y} = \text{sign}(f(\mathbf{x}))$. Later in Fig. 5(a), we can see the classes (see the green ellipses as contours) and the optimal boundary (see the red solid curve).

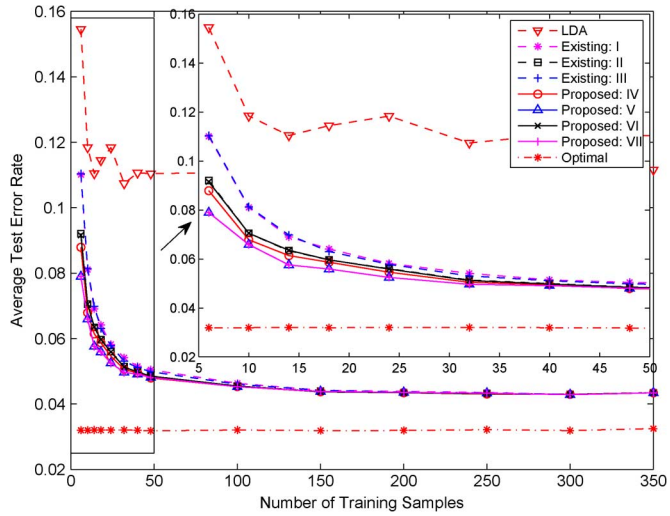

Fig. 4. Synthetic example with different covariance matrices: average test error rate versus number of training samples.

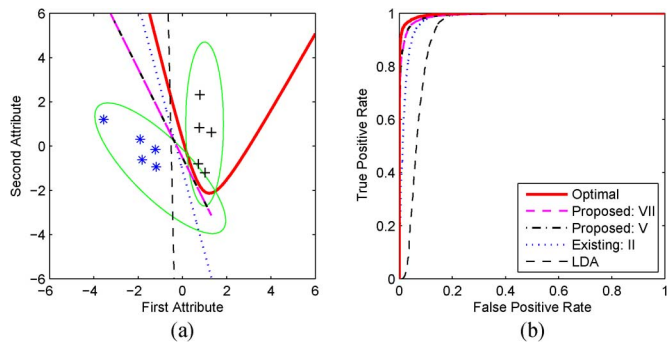

Fig. 5. One specific synthetic realization with different covariance matrices: existing SVM II may be misleading when there are not enough training samples (blue dotted line), this drawback can be overcome by the proposed methods (black dash-dotted line and magenta dashed line). (a) Data and classifiers. (b) ROC curves.

The simulation setup is the same as that for the previous synthetic experiment. Fig. 4 shows the results of the average test error rate versus the number of training samples. Similar to Fig. 1, we can see that the proposed SVMs outperform the existing ones, especially when the training sample size is small. However, there are also two differences when comparing Fig. 4 with Fig. 1. First, the LDA performs relatively worse than the SVMs and is not stably decreasing as N goes larger. This is because that the theoretical LDA method is no longer optimal and it does not take the classification error into problem formulation. Second, there always exists gap between the linear SVMs and the optimal boundary. Again, this is because the optimal boundary is quadratic however the linear SVMs can only provide linear classifiers.

Fig. 5 shows one specific realization. We can see that the proposed methods are better than the existing ones and the LDA based on the ROC curves in Fig. 5(b).

Fig. 6 shows the average test error rate versus the tuning parameter C for $N = 6, 10, 18, 48$. We observe the first three observations similar to that from Fig. 3, however, now we cannot

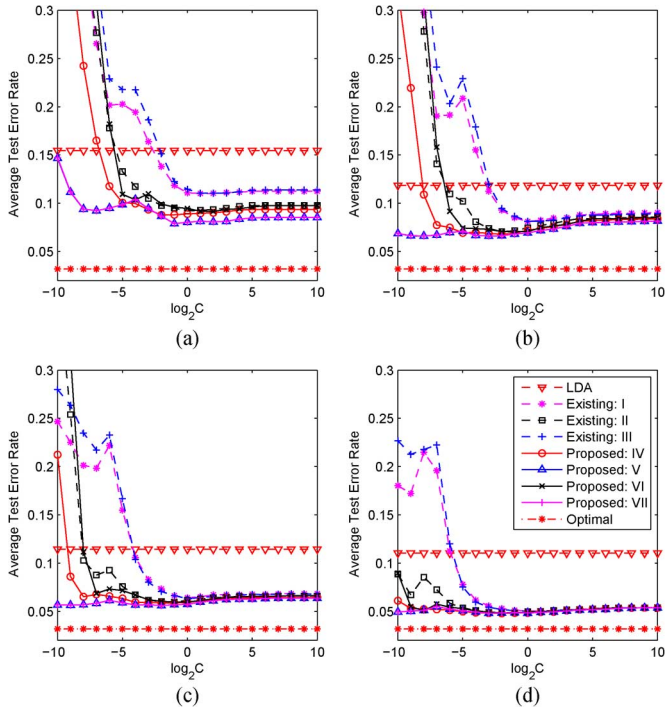

Fig. 6. Synthetic example with different covariance matrices: average test error rate versus tuning parameter C. (a) N = 6. (b) N = 10. (c) N = 18. (d) N = 48.

TABLE III
SUMMARY ON DATA SETS FROM DIFFERENT SOURCES

No.	Data Set	d	Set Size	Source	# Realz.
1	fourclass	2	862	[27]	120
2	liver disorders	6	345	UCI	80
3	pima indians diabetes	8	768	UCI	80
4	heart	13	270	Statlog	80
5	german	24	1000	Statlog	80
6	ionosphere	34	351	UCI	80
7	sonar	60	208	UCI	120
8	USPS	256	2200	UCI	20
9	MNIST	784	11791/1991 (train/test)	[28]	10
10	CMU face	960	312	UCI	5

observe that the empirical LDA performs more and more closely to the optimal one as N becomes larger, which coincides with the results in Fig. 4.

C. Experiments on Real Data

Now we move to the experiments on real data. Table III briefly summarizes different real data sets used in this paper.

Real Data Sets: For the real experiments, we consider different binary classification data sets from different sources, i.e., Statlog [29] and UCI [27], [30], and [28], etc. The data dimensions vary from less than ten (i.e., No. 1–3) to less than one hundred (i.e., No. 4–7), and then to several hundreds (i.e., No. 8–10). The sizes of the data sets are generally around several hundreds (i.e., No. 1–7 and 10) or even much more (i.e., No. 8 and 9). Such various data sets provide enough different real feature label distributions for the proposed and existing SVMs to explore. To avoid the features with larger dynamic range dominating those with smaller ones and the numerical difficulties during the calculation [13], [14], for data sets No. 1–7, we adopt the scaled version of the different real data sets available[6] in the popular SVM package LIBSVM [31] and the data set fourclass is transformed to two-class in LIBSVM. For the grayscale image data sets, i.e., data sets No. 8–10, we linearly transform the pixel values from [0, 255] to [−1, 1]. For the digits classification data sets USPS and MNIST, we focus on the most difficult task, e.g., the classification between the similar digits four and nine. For the face recognition data set CMU face, we consider the classification between "look at left" and "look at right".

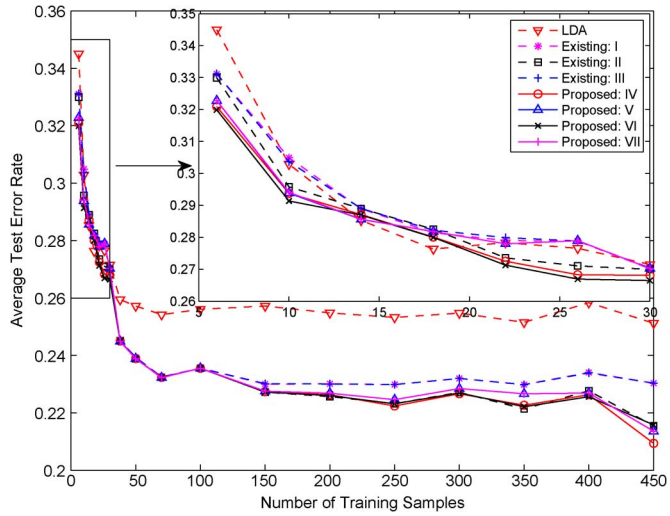

Fig. 7. Fourclass: average test error rate versus number of training samples.

Experiment Setup: For each realization, similar to the reason in the synthetic experiments, we first randomly select equal number of training samples for both classes when N is relatively small compared with data dimensions, that is, when $N \leq 30$ for fourclass in Fig. 7 or N takes the values for other data sets in Fig. 10. When $N > 30$ for fourclass data set, we randomly select N samples from the whole data set and leave the remaining ones for testing. For the MNIST data set, the training samples are drawn from the training set and the whole test set is always used for testing. We repeat the realization for each data set. The number of realizations for each real data set is in column "# Realz." in Table III. The other parameter settings, e.g., the values of C, ν, ν_1 and ν_2, are the same as that for the synthetic experiments.

We begin with the real visualizable data set fourclass since it has only two features. Fig. 7 shows the results of the average test error rate versus the number of training samples. We have that the proposed methods outperform the existing ones when the number of training samples is relatively small, e.g., $N \leq 30$. The differences between the average test error rates are becoming smaller as the number of training samples is getting larger, e.g., $30 < N \leq 100$. This is quite similar to what we have observed in the synthetic numerical experiments. When N is even larger, $N > 100$, it seems ℓ_2-norm SVMs perform relatively better. This may mean that the feature label distribution is also important. Here, as before, we are more interested in the

[6]http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html

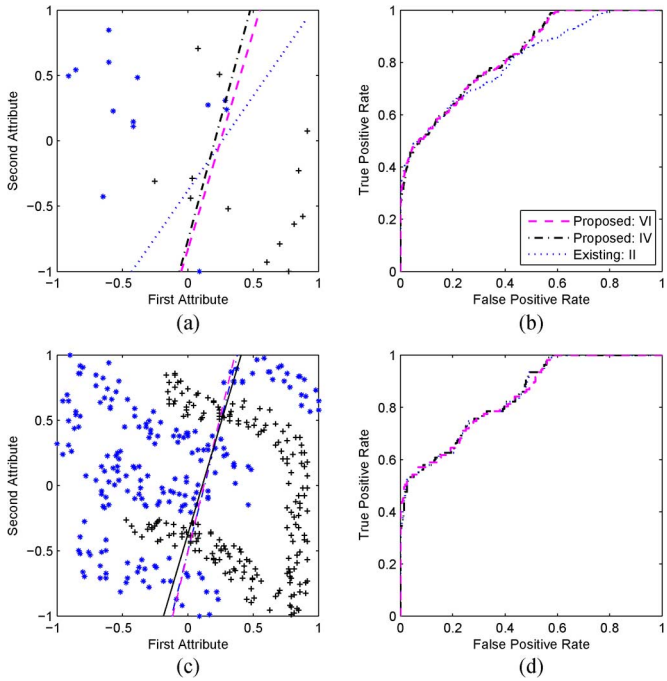

Fig. 8. Fourclass: small versus large number of training sizes. (a) N = 26. (b) ROC curves. (c) N = 400. (d) ROC curves.

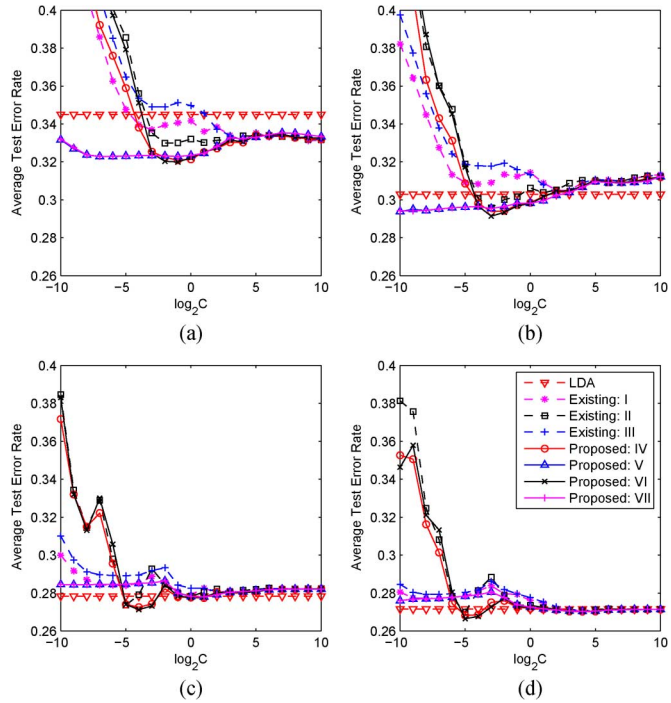

Fig. 9. Fourclass: average test error rate versus tuning parameter C. (a) N = 6. (b) N = 10. (c) N = 22. (d) N = 30.

regime that the number of samples is relatively not too large, say $N \leq 30$ for `fourclass`. Comparing Fig. 7 with Figs. 1 and 4, we can observe similar but more nonsmooth decay trend. This is reasonable because that the numerical results depend strongly on the feature label distribution [12] and the real data `fourclass` has much more complicated feature label distribution than the synthetic data sets.

Similar to Fig. 2, Fig. 8 shows two specific realizations. Based on the ROC curves, we see that when N is small, the proposed methods provide better linear classifiers, and when N is large, all the methods provide similar linear classifiers.

Fig. 9 shows the average test error rate versus the tuning parameter C. It shows similar patterns as Figs. 3 and 6.

For other real data sets, similar results can be obtained and for clarity of presentation and due to space limitations, we only focus on the scenarios that the training sets relatively small compared to the data dimensions (e.g., see the values of N in Fig. 10 for different data sets).

Fig. 10 shows all the numerical results. First of all, it is obvious that the LDA is much more worse than SVMs. Figs. 10(a), 10(e), and 10(i) show that the proposed methods perform better than the existing ones when the number of training samples N is very small, say $N = 10, 14, 18, 22, 26$ for `liver disorders`, $N = 10, 16, 22, 28$ for `ionosphere` and $N = 20, 60, 100, 150$ for `CMU face`. However, when N gets relatively larger for `liver disorders` (e.g., $N = 30, 34$) and `CMU face` (e.g., $N = 200$) all the methods perform really closely to each other. When N gets larger for `ionosphere` (e.g., $N \geq 34$), the existing method II slightly outperforms the proposed ones, however, the proposed methods IV and VI still perform quite closely to the best method II.

Figs. 10(b), 10(c) and 10(d) look similar. When N is too small, e.g., $N = 12, 16$ for `pima indians diabetes`, $N = 22, 26, 30$ for `heart`, and $N = 10, 22$ for `german`, the proposed methods (mainly methods IV and VI) and the existing method II perform closely to each other. This may due to that N is too small for the data set and the advantage of the proposed methods is not so obvious. Once N gets slightly larger, e.g., $N \geq 20$ for `pima indians diabetes`, $N \geq 34$ for `heart` and $N \geq 34$ for `german`, the proposed methods obviously perform better than the existing method II. Meanwhile, the differences between the proposed methods and the other existing methods I and III become smaller.

Fig. 10(f) shows the results for `sonar`. Since the data set `sonar` has 60 attributes and there are only 208 samples in total, we cannot set the number of training samples to be too large. Fortunately, for almost all the number of training samples we have simulated, the proposed methods (especially method VI) always tend to outperform the existing ones.

Figs. 10(g) and 10(h) look similar. We can see that all the proposed methods outperform or at least are comparable to the existing methods for all the simulated number of samples.

To summarize, even though the simulation results based on different synthetic and real data sets are different from each other, however, they do share a similar pattern: when the number of training samples N is not large relative to the data dimension, the proposed methods tend to perform better than the existing ones. The fact that all the twelve numerical experiments (e.g., two synthetic and ten real experiments) we have simulated share the similar pattern strongly convinces us that the proposed methods based on our proposed unified framework do perform better when the number of training samples is small (e.g., see Figs. 1, 4, 7, and 10), and we also have nice interpretations and insights (for instance, see Figs. 2, 5, and 8). Thus, the proposed unified framework does benefit us with more optional SVMs.

VIII. Conclusion

In this paper, we have proposed a unified framework that can characterize both the proposed and many existing SVMs

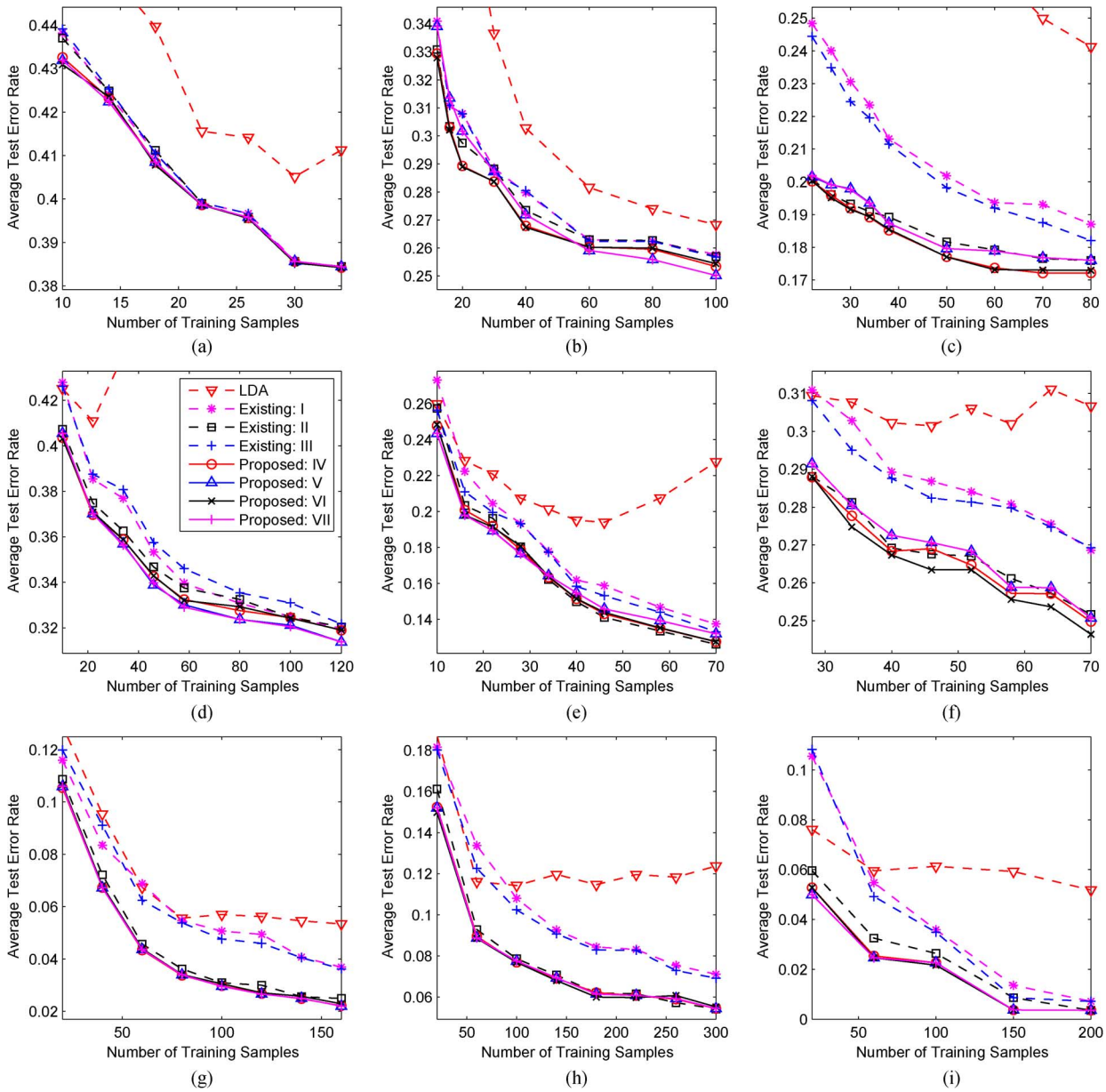

Fig. 10. Average test error rate versus number of training samples for different real data sets. (a) liver. (b) pima. (c) heart. (d) german. (e) ionosphere. (f) sonar. (g) USPS. (h) MNIST. (i) CMU face.

by simply selecting different types of weighted vector norms. The origin of having such different methods is based on the general linear SVM problem formulation with different normalization measures. The unified framework can provide us with more insights and help us understand the connections and differences between different SVMs. The numerical experiments on both the synthetic and real data sets show that the proposed methods derived from the unified framework outperform the existing ones when the number of training samples is not large.

Appendix
Proofs and Derivations

A. Proof of Proposition 1

For any point $(\boldsymbol{\beta}, \beta_0, \mathbf{F}, \boldsymbol{\xi}, t)$ feasible for problem (14), it is also feasible for problem (15), thus $v^\star((14)) \geq v^\star((15))$.

Now, since $(\boldsymbol{\beta}_2^\star, \beta_{02}^\star, \mathbf{F}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is optimal for problem (15), and we can always project $\boldsymbol{\beta}_2^\star$ orthogonally onto two orthogonal subspace $\mathcal{R}(\mathbf{F}_2^\star)$ and $\mathcal{N}(\mathbf{F}_2^{\star\top})$: $\boldsymbol{\beta}_2^\star = \mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star + (\mathbf{I} - \mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)})\boldsymbol{\beta}_2^\star$. It is easy to verify that $(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star, \beta_{02}^\star, \mathbf{F}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is feasible and therefore optimal for problem (15). Furthermore, we have

$$\begin{aligned}
&\mathbf{F}_2^{\star\top}\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)^\top \mathbf{F}_2^\star \preceq t_2^\star \mathbf{F}_2^{\star\top}\mathbf{F}_2^\star \\
&\quad\Longleftrightarrow \mathbf{x}^\top\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)^\top \mathbf{x} \\
&\quad\leq t_2^\star \mathbf{x}^\top\mathbf{x},\ \forall \mathbf{x} \in \mathcal{R}\left(\mathbf{F}_2^\star\right) \\
&\quad\Longleftrightarrow \mathbf{x}^\top\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)^\top \mathbf{x} \leq t_2^\star \mathbf{x}^\top\mathbf{x},\ \forall \mathbf{x} \\
&\quad\Longleftrightarrow \left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)\left(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star\right)^\top \preceq t_2^\star \mathbf{I}. \qquad (68)
\end{aligned}$$

Then, it is obvious that $(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star, \beta_{02}^\star, \mathbf{F}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is also feasible for problem (14). Recall the relationship

$v^\star((14)) \geq v^\star((15))$, we then have $v^\star((14)) = v^\star((15))$, and $(\mathbf{P}_{\mathcal{R}(\mathbf{F}_2^\star)}\boldsymbol{\beta}_2^\star, \beta_{02}^\star, \mathbf{F}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is optimal for problem (14).

As pointed out before, the optimal solution $(\boldsymbol{\beta}_1^\star, \beta_{01}^\star, \mathbf{F}_1^\star, \boldsymbol{\xi}_1^\star, t_1^\star)$ for problem (14) must also be feasible for problem (15), then considering $v^\star((14)) = v^\star((15))$, it must also be optimal for problem (15).

B. Proof of Proposition 2

The SDR of problem (16) is:

$$\begin{aligned} \underset{\mathbf{v},\beta_0,\mathbf{T},\boldsymbol{\xi},t}{\text{minimize}} \quad & \frac{1}{2}t + C\mathbf{1}^\top\boldsymbol{\xi} \\ \text{subject to} \quad & y_i(\mathbf{v}^\top\mathbf{x}_i + \beta_0) \geq 1 - \xi_i \quad \forall i, \\ & \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \{\mathrm{Tr}(\mathbf{T}\mathbf{A}_{\boldsymbol{\theta}})\} \leq 1, \\ & \begin{bmatrix} t & \mathbf{v}^\top \\ \mathbf{v} & \mathbf{T} \end{bmatrix} \succeq \mathbf{0}, \\ & \boldsymbol{\xi} \geq \mathbf{0}. \end{aligned} \tag{69}$$

For any point $(\mathbf{v}_1, \beta_{01}, \mathbf{T}_1, \boldsymbol{\xi}_1, t_1)$ feasible for problem (69), we can readily see that $(\mathbf{v}_1, \beta_{01}, \boldsymbol{\xi}_1)$ is feasible for problem (17), and we have for $t_1 > 0$:

$$\begin{aligned} & \frac{1}{t_1}\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left\| \mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}_1 \right\|_2^2 \right\} \leq \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{Tr}\left(\mathbf{A}\mathbf{T}_1\right) \leq 1 \\ \Longrightarrow & \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left\| \mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}_1 \right\|_2^2 \right\} \leq t_1. \end{aligned} \tag{70}$$

As for $t_1 = 0$, we must have $\mathbf{v}_1 = \mathbf{0}$ according to the linear matrix inequality constraint in problem (69) and (70) still holds. That is, we always have $v^\star((17)) \leq v^\star((69))$.

Now, assume $(\mathbf{v}_2^\star, \beta_{02}^\star, \boldsymbol{\xi}_2^\star)$ is optimal for (17), we can always construct $t_2^\star = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\|\mathbf{A}_{\boldsymbol{\theta}}^{1/2}\mathbf{v}_2^\star\|_2^2\}$. Note that under the condition that there exists some $\mathbf{A}_{\boldsymbol{\theta}}$ is full rank, if $t_2^\star = 0$, we must have $\mathbf{v}_2^\star = \mathbf{0}$. Thus, we can choose $\mathbf{T}_2^\star$ as: $\mathbf{T}_2^\star = \begin{cases} \mathbf{0} & t_2^\star = 0 \\ \frac{\mathbf{v}_2^\star\mathbf{v}_2^{\star\top}}{t_2^\star} & t_2^\star > 0. \end{cases}$ Then, obviously $(\mathbf{v}_2^\star, \beta_{02}^\star, \mathbf{T}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is feasible for problem (69) with the objective equal to that in problem (17). Together with $v^\star((17)) \leq v^\star((69))$, straightforwardly we can conclude that $v^\star((17)) = v^\star((69))$, and thus $(\mathbf{v}_2^\star, \beta_{02}^\star, \mathbf{T}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is also optimal for problem (69). Note that we also have $\mathrm{rank}(\mathbf{T}_2^\star) \leq \min(m, d)$ satisfied, thus the SDR (69) equals to the SDP (16).

C. Proof of Proposition 3

Given any point $(\mathbf{v}_1, \beta_{01}, \mathbf{T}_1, \boldsymbol{\xi}_1, t_1)$ feasible for problem (18), if for some i, $\mathbf{T}_{1ii} = 0$, we must have $\mathbf{v}_{1i} = 0$, otherwise the point cannot be feasible due to the linear matrix inequality and $\mathbf{T}$ being diagonal constraints. For simplicity of presentation, we make the convention that $\mathbf{v}_{1i}^2/\mathbf{T}_{1ii} = 0$ when $\mathbf{T}_{1ii} = 0$. Then we can readily see that $(\mathbf{v}_1, \beta_{01}, \boldsymbol{\xi}_1)$ is feasible for problem (19), and we have

$$\begin{aligned} t_1 &\geq \sum_{i=1}^n \mathbf{v}_{1i}^2/\mathbf{T}_{1ii} \geq \left(\sum_{i=1}^n \mathbf{v}_{1i}^2/\mathbf{T}_{1ii}\right) \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \sum_{i=1}^n \mathbf{T}_{1ii}\mathbf{A}_{\boldsymbol{\theta}ii} \right\} \\ &= \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left(\sum_{i=1}^n \mathbf{v}_{1i}^2/\mathbf{T}_{1ii}\right) \left(\sum_{i=1}^n \mathbf{T}_{1ii}\mathbf{A}_{\boldsymbol{\theta}ii}\right) \right\} \\ &\geq \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left(\sum_{i=1}^n \sqrt{\mathbf{T}_{1ii}\mathbf{A}_{\boldsymbol{\theta}ii}} \cdot \sqrt{\mathbf{v}_{1i}^2/\mathbf{T}_{1ii}}\right)^2 \right\} \\ &= \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left(\sum_{i=1}^n \sqrt{\mathbf{A}_{\boldsymbol{\theta}ii}}\,|\mathbf{v}_{1i}|\right)^2 \right\} \\ &= \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left\| \mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})\mathbf{v}_1 \right\|_1^2 \right\} \end{aligned} \tag{71}$$

where the third inequality "$\geq$" is due to Cauchy-Schwartz inequality property and the equality holds if and only if $\mathbf{T}_{1ii}\mathbf{A}_{\boldsymbol{\theta}ii}$ and $\mathbf{v}_{1i}^2/\mathbf{T}_{1ii}, \forall i = 1, \ldots, n$, have the same linear dependence, that is, there exists some α such that

$$\alpha = \frac{\mathbf{v}_{1i}^2/\mathbf{T}_{1ii}}{\mathbf{T}_{1ii}\mathbf{A}_{\boldsymbol{\theta}ii}} = \frac{\mathbf{v}_{1i}^2}{\mathbf{T}_{1ii}^2\mathbf{A}_{\boldsymbol{\theta}ii}}, \quad \forall i = 1, \ldots, n. \tag{72}$$

According to (71), we can conclude that $v^\star((19)) \leq v^\star((18))$

Now, assume $(\mathbf{v}_2^\star, \beta_{02}^\star, \boldsymbol{\xi}_2^\star)$ is optimal for (19). Under the condition that there exists some full rank $\mathrm{Diag}(\mathbf{A}_{\boldsymbol{\theta}})$, we have:

- If $\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\|\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})\mathbf{v}_2^\star\|_1^2\} = 0$, then $\mathbf{v}_2^\star = \mathbf{0}$, and we choose $\mathbf{T}_2^\star = \mathbf{0}$ and $t_2^\star = 0$.
- If $\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\|\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})\mathbf{v}_2^\star\|_1^2\} > 0$, we can first find some $\mathbf{A}_{\boldsymbol{\theta}^\star}$ such that $\boldsymbol{\theta}^\star = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\|\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})\mathbf{v}_2^\star\|_1^2\}$. Then inspired by (72), we can construct the diagonal matrix $\mathbf{T}_2^\star$ such that $\mathbf{T}_{2ii}^{\star 2} = \mathbf{v}_{2i}^{\star 2}/(\alpha\mathbf{A}_{\boldsymbol{\theta}^\star ii})$, e.g., $\mathbf{T}_{2ii}^\star = |\mathbf{v}_{2i}^\star|/\sqrt{\alpha\mathbf{A}_{\boldsymbol{\theta}^\star ii}}$, where α is given by solving $\mathrm{Tr}(\mathbf{T}_2^\star\mathbf{A}_{\boldsymbol{\theta}^\star}) = 1$, that is, $\alpha = \|\mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}^\star})\mathbf{v}_2^\star\|_1^2$. After that, we furthermore construct $t_2^\star = \sum_{i=1}^n \mathbf{v}_{2i}^{\star 2}/\mathbf{T}_{2ii}^\star$ and get

$$\begin{aligned} t_2^\star &= \sum_{i=1}^n \mathbf{v}_{2i}^{\star 2}/\mathbf{T}_{2ii}^\star = \sum_{i=1}^n \mathbf{v}_{2i}^{\star 2}\frac{\sqrt{\alpha\mathbf{A}_{\boldsymbol{\theta}^\star ii}}}{|\mathbf{v}_{2i}^\star|} \\ &= \left(\sum_{i=1}^n |\mathbf{v}_{2i}^\star|\sqrt{\mathbf{A}_{\boldsymbol{\theta}^\star ii}}\right) \left\| \mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\theta}^\star}\right)\mathbf{v}_2^\star \right\|_1 \\ &= \left\| \mathrm{Diag}^{1/2}\left(\mathbf{A}_{\boldsymbol{\theta}^\star}\right)\mathbf{v}_2^\star \right\|_1^2 \\ &= \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\{ \left\| \mathrm{Diag}^{1/2}(\mathbf{A}_{\boldsymbol{\theta}})\mathbf{v}_2^\star \right\|_1^2 \right\}. \end{aligned} \tag{73}$$

Therefore, no matter what the case is, we always have that $(\mathbf{v}_2^\star, \beta_{02}^\star, \mathbf{T}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is feasible for problem (18) with the objective equal to that in problem (19). Combining with $v^\star((19)) \leq v^\star((18))$, we can have $v^\star((19)) = v^\star((18))$, and thus $(\mathbf{v}_2^\star, \beta_{02}^\star, \mathbf{T}_2^\star, \boldsymbol{\xi}_2^\star, t_2^\star)$ is also optimal for problem (18).

REFERENCES

[1] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 15th Annu. Workshop Comput. Learn. Theory*, Pittsburgh, PA, USA, Jul. 1992, pp. 144–152.

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

[3] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[4] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2000.

[5] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

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

[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[8] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani, "1-norm support vector machines," in *Adv. Neural Inf. Process. Syst.*, 2004, pp. 49–56.

[9] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th Int. Conf. Mach. Learn.*, Madison, WI, USA, Jun. 1998, pp. 82–90.

[10] A. B. Chan, V. Nuno, and G. R. Lanckriet, "Direct convex relaxations of sparse SVM," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvallis, OR, USA, June 2007, pp. 145–153.
[11] M. Tan, L. Wang, and I. W. Tsang, "Learning sparse SVM for feature selection on very high dimensional datasets," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, Jun. 2010, pp. 1047–1054.
[12] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformat.*, vol. 21, no. 8, pp. 1509–1515, 2005.
[13] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2006.
[14] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, Applications*. Philadelphia, PA, USA: SIAM, 2007, vol. 20.
[15] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," Tech. Rep., 2003.
[16] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 131–159, 2002.
[17] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Inf. Process. Syst.*, 2000, pp. 668–674.
[18] Y. Grandvalet and S. Canu, "Adaptive scaling for feature selection in SVMs," in *Adv. Neural Inf. Process. Syst.*, 2002, pp. 553–560.
[19] M. H. Nguyen and F. De la Torre, "Optimal feature selection for support vector machines," *Pattern Recogn.*, vol. 43, no. 3, pp. 584–591, 2010.
[20] Y. Tian, Y. Shi, and X. Liu, "Recent advances on support vector machines research," *Technol. Econom. Develop. Econ.*, vol. 18, no. 1, pp. 5–33, 2012.
[21] D. P. Palomar and Y. Jiang, "MIMO transceiver design via majorization theory," *Found. Trends Commun. Inf. Theory*, vol. 3, no. 4, pp. 331–551, 2006.
[22] 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.
[23] D. Goldfarb and G. Iyengar, "Robust portfolio selection problems," *Math. Oper. Res.*, vol. 28, no. 1, pp. 1–38, 2003.
[24] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[25] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[26] MOSEK, The MOSEK Optimization Toolbox for MATLAB Manual, 2013 [Online]. Available: http://www.mosek.com, Tech. Rep.
[27] T. K. Ho and E. M. Kleinberg, "Building projectable classifiers of arbitrary complexity," in *Proc. 13th IEEE Int. Conf. Pattern Recogn.*, 1996, vol. 2, pp. 880–885, .
[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[29] R. D. King, C. Feng, and A. Sutherland, "Statlog: Comparison of classification algorithms on large real-world problems," *Appl. Artif. Intell. Int. J.*, vol. 9, no. 3, pp. 289–333, 1995.
[30] K. Bache and M. Lichman, UCI Machine Learning Repository, 2013 [Online]. Available: http://archive.ics.uci.edu/ml
[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.* vol. 2, pp. 27:1–27:27, 2011 [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/libsvm

Yiyong Feng received a B.E. degree in Electronic and Information Engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010. Since then he has been pursuing a Ph.D. degree in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST). From March 2013 to August 2013, Mr. Feng was with the Systematic Market-Making Group at Credit Suisse (Hong Kong). His research interests are in convex optimization, nonlinear programming, and robust optimization, with applications in signal processing, financial engineering, and machine learning.

Daniel P. Palomar (S'99–M'03–SM'08–F'12) received the Electrical Engineering and Ph.D. degrees from the Technical University of Catalonia (UPC), Barcelona, Spain, in 1998 and 2003, respectively.

He is a Professor in the Department of Electronic and Computer Engineering at the Hong Kong University of Science and Technology (HKUST), Hong Kong, which he joined in 2006. Since 2013 he is a Fellow of the Institute for Advance Study (IAS) at HKUST. He had previously held several research appointments, namely, at King's College London (KCL), London, U.K.; Technical University of Catalonia (UPC), Barcelona; Stanford University, Stanford, CA; Telecommunications Technological Center of Catalonia (CTTC), Barcelona; Royal Institute of Technology (KTH), Stockholm, Sweden; University of Rome "La Sapienza", Rome, Italy; and Princeton University, Princeton, NJ. His current research interests include applications of convex optimization theory, game theory, and variational inequality theory to financial systems and communication systems.

Dr. Palomar is a recipient of a 2004/06 Fulbright Research Fellowship, the 2004 Young Author Best Paper Award by the IEEE Signal Processing Society, the 2002–2003 best Ph.D. prize in information technologies and communications by the Technical University of Catalonia (UPC), the 2002–2003 Rosina Ribalta first prize for the Best Doctoral Thesis in information technologies and communications by the Epson Foundation, and the 2004 prize for the best Doctoral Thesis in Advanced Mobile Communications by the Vodafone Foundation and COIT.

He has been an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY and of the IEEE TRANSACTIONS ON SIGNAL PROCESSING, a Guest Editor of the IEEE SIGNAL PROCESSING MAGAZINE 2010 Special Issue on Convex Optimization for Signal Processing, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2008 Special Issue on Game Theory in Communication Systems, and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 2007 Special Issue on Optimization of MIMO Transceivers for Realistic Communication Networks.