Buffered Probability Minimization: A New Interpretation of Robust and Regularized Support Vector Machines

November 12, 2016

Matthew Norton, *mdnorto@gmail.com*, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL.
Stan Uryasev, *uryasev@ufl.edu*, Director of Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL .

Abstract

We approach the problem of binary classification by minimizing Buffered Probability of Exceedance (bPOE), a new characterization of uncertainty proved to be the minimal quasiconvex upper bound of Probability of Exceedance (POE), i.e. the expected value of the 0-1 loss. We prove that a broad class of regularized and robust Support Vector Machine (SVM) formulations are special cases of this approach. This provides a new statistical perspective on the optimality of the hinge loss as an approximation of 0-1 loss, the meaning of the SVM's free parameters, the meaning of its objective function, and the interpretation of the classification margin induced by the choice of regularizer. Additionally, it provides a connection between regularization and robustness that extends the current views in the literature. Overall, we find that bPOE minimization provides a unifying view of many SVM variants, including regularized, robust, convex, and non-convex formulations.

1 Introduction

For the task of binary classification, SVM's (Cortes and Vapnik (1995)) have proven to be an extremely popular tool for classification. With SVM's based upon sound geometric intuition and statistical learning principals, there have been many proposed customizations of the base formulation to deal with special circumstances such as imbalance classes (Osuna et al. (1997)), cost sensitivity, semi-supervised and manifold learning (Belkin et al. (2006), Shen et al. (2015)), and AUC maximization (Herbrich et al. (1999), Rakotomamonjy (2004), Brefeld and Scheffer (2005)). In addition, to improve generalization, designs have been proposed that use different regularizers (Wang et al. (2006)) as well as ideas from robust optimization (Katsumata and Takeda (2015), Trafalis and Gilbert (2007), Bhattacharyya et al. (2005), El Ghaoui et al. (2003)) where data is

viewed as lying in some fixed uncertainty set. Nevertheless, two primary components are relevant in most formulations. First, the hinge loss function is used as a convex upper bound to the 0-1 loss. Second, regularization or robust optimization is introduced for improved generalization.

In this paper, we approach the binary classification problem from a statistical perspective by minimizing bPOE, a new characterization of uncertainty proved to be the minimal quasiconvex upper bound of POE, i.e. the expected value of the 0-1 loss. We prove that a broad class of regularized and robust SVM formulations are special cases of this approach, particularly its convex case. This provides a new statistical perspective on the optimality of the hinge loss as an approximation of 0-1 loss, the meaning of the SVM's free parameters, the meaning of its objective function, and the interpretation of the classification margin induced by the choice of regularizer.

Additionally, this new perspective allows us to connect regularization and robustness, extending the current views in the literature. In Xu et al. (2009), Katsumata and Takeda (2015), SVM's regularized with a norm, or with convex combinations of norms, are shown to be equivalent to robust formulations with uncertainty sets characterized by the associated dual norm. We generalize this view and show that a broader class of regularized SVM's can be posed as robust SVM's where the choice of uncertainty set is determined by the regularization function and free parameter. Specifically, we show that the uncertainty set is the convex set which has support function equal to the regularization function scaled by the free parameter.

Finally, after showing that many convex SVM's are special cases of the convex case of the bPOE minimization problem, we show that the bPOE minimization also has a non-convex case. Thus, we can view many cases of the bPOE minimization as being an extended formulation of the related SVM's, where the range of free parameter is extended to provide a wider variety of potentially optimal hyperplanes. We show that some of the existing non-convex SVM's in the literature are also special cases of this bPOE minimization problem.

The remainder of this paper is organized as follows. In Section 2, we review bPOE and the concept of the superquantile, along with its properties and calculation formula. In Section 3, we first review a common form of regularized SVM's. Then, we pose a bPOE minimization problem for binary classification and show that for a certain parameter range, the regularized SVM and bPOE

minimization problem provide the same set of optimal solutions. We then show that this provides new interpretations for aspects of the SVM. In Section 4, we consider a robust bPOE minimization problem, show that the regularized SVM's form a special case, and draw connections between the choice of regularizer and the equivalent choice of uncertainty set. In Section 5, we briefly address the non-convex case of the bPOE minimization problem, show how it fits into the current SVM literature, and provide an intuitive interpretation from a robust optimization perspective. We conclude in Section 6. Appendix A overviews the examples given in Table 1 and Table 2 of SVM's that are special cases of bPOE minimization, providing also references and notational information.

2 Background: Buffered Probability of Exceedance and Superquantiles

When working with optimization of tail probabilities, one frequently works with constraints or objectives involving *probability of exceedance* (POE), $p_z(Z) = P(Z > z)$, or its associated quantile $q_\alpha(Z) = \min\{z | P(Z \leq z) \geq \alpha\}$, where $\alpha \in [0, 1]$ is a probability level and $z \in \mathbb{R}$ is a fixed threshold level. The quantile is a popular measure of tail probabilities in financial engineering, called within this field Value-at-Risk by its interpretation as a measure of tail risk. The quantile, though, when included in optimization problems via constraints or objectives, is quite difficult to treat with continuous (linear or non-linear) optimization techniques.

A significant advancement was made by Rockafellar and Uryasev Rockafellar and Uryasev (2000) in the development of an approach to combat the difficulties raised by the use of the quantile function in optimization. They explored a replacement for the quantile, called CVaR within the financial literature, and called the superquantile in a general context. The superquantile is a measure of uncertainty similar to the quantile, but with superior mathematical properties. Formally, the superquantile (CVaR) for a continuously distributed real valued random variable Z is defined as

$$\bar{q}_\alpha(Z) = E\left[Z | Z > q_\alpha(Z)\right]. \tag{1}$$

For general distributions, the superquantile can be defined by the following formula,

$$\bar{q}_\alpha(Z) = \min_\gamma \quad \gamma + \frac{E[Z-\gamma]^+}{1-\alpha}, \tag{2}$$

where $[\cdot]^+ = \max\{\cdot, 0\}$. Similar to $q_\alpha(Z)$, the superquantile can be used to assess the tail of the distribution but is far easier to handle in optimization contexts. It also has the important property that it considers the magnitude of events within the tail.

Working to extend this concept, bPOE was developed as the inverse of the superquantile in the same way that POE is the inverse of the quantile[1]. bPOE is defined in the following way, where $\sup Z$ denotes the essential supremum of random variable Z.

Definition 1. *bPOE for a random variable Z at a threshold z equals*

$$\bar{p}_z(Z) = \begin{cases} \max\{1-\alpha | \bar{q}_\alpha(Z) \geq z\}, & \text{if } z \leq \sup Z, \\ 0, & \text{otherwise.} \end{cases}$$

In words, for any threshold $z \in (E[Z], \sup Z)$, bPOE can be interpreted as one minus the probability level at which the tail expectation, or superquantile, equals z. Although bPOE seems troublesome to calculate, Norton and Uryasev (2014) provides the following calculation formula for bPOE.

Proposition 1. *Given a real valued random variable Z and a fixed threshold z, bPOE for random variable Z at z equals*

$$\bar{p}_z(Z) = \inf_{\gamma<z} \frac{E[Z-\gamma]^+}{z-\gamma} = \begin{cases} \lim\limits_{\gamma\to-\infty} \ \frac{E[Z-\gamma]^+}{z-\gamma} = 1, & \text{if } z \leq E[Z], \\ \min\limits_{\gamma<z} \ \frac{E[Z-\gamma]^+}{z-\gamma}, & \text{if } z \in (E[Z], \sup Z), \\ \lim\limits_{\gamma\to z^-} \ \frac{E[Z-\gamma]^+}{z-\gamma} = P(Z = \sup Z), & \text{if } z = \sup Z, \\ \min\limits_{\gamma<z} \ \frac{E[Z-\gamma]^+}{z-\gamma} = 0, & \text{if } \sup Z < z. \end{cases} \tag{3}$$

It is also important to note that formula (3) has the following properties.

Property 1 (Norton and Uryasev (2014))**.** *If $z \in (E[Z], \sup Z)$ and* $\min\limits_{\gamma<z} \frac{E[Z-\gamma]^+}{z-\gamma} = \frac{E[Z-\gamma^*]^+}{z-\gamma^*} =$

[1]bPOE was also developed as a generalization to Buffered Probability of Failure from Rockafellar and Royset (2010)

$1-\alpha^*$, *then:*

$$\bar{p}_z(Z) = 1-\alpha^*, \qquad \bar{q}_{\alpha^*}(Z) = z, \qquad q_{\alpha^*}(Z) = \gamma^*.$$

Property 2 (Mafusalov and Uryasev (2015))**.** *bPOE is the minimal quasiconvex upper bound of POE.*

Thus, using formula (3), bPOE can be efficiently calculated and we can recover quantile and superquantile information. Furthermore, bPOE has many advantageous mathematical properties, which Mafusalov and Uryasev (2015) explores in depth, and is shown to be the best quasiconvex upper bound for POE.

3 SVM's and bPOE Minimization

3.1 Regularized SVM's

In the typical machine learning setup for binary classification, we have a random feature vector $X \in \mathbb{R}^n$, random label $Y \in \{-1, +1\}$, and N observations of random pairs $(X_i, Y_i), i = 1, ..., N$. We then would like to use these observations to learn a scoring function $h : \mathbb{R}^N \to \mathbb{R}$ to properly predict the class of unlabelled feature vectors via a decision function $Y = sign(h(X))$. In the context of SVM's, the method of finding the best classifier is to use the scoring function $h(w, b, X) = w^T X + b$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$ and to minimize the probability of misclassifications, which is written in terms of the expected value of the 0-1 loss function. Specifically, with 0-1 loss written as the indicator function $I_{\{-Y(w^T X+b)>0\}}$, we have that the probability of misclassification is the expected value of the 0-1 loss,

$$P(-Y(w^T X + b) > 0) = E\left[I_{\{-Y(w^T X+b)>0\}}\right] .$$

Since the 0-1 loss is difficult to handle in optimization and also does not take into account the magnitude of misclassification errors, SVM's approximate this objective by using a convex upper bound of the 0-1 loss, specifically the hinge loss $[-Y(w^T X + b) + 1]^+$, and minimize its expectation,

$$\min_{w,b} E\left[-Y(w^T X + b) + 1\right]^+ .$$

In addition, a regularization term $g : \mathbb{R}^n \to \mathbb{R}$ is added to this objective function with a tradeoff parameter $z \leq 0$ to produce a classifier with a large margin[2] or to simply penalize the complexity of the resulting classifier. They then attempt to solve a problem of the following form, where $g(w)$ is typically non-negative, convex, and positive homogenous (PH)[3] and we have parameters $p_i \in [0, 1]$, which are probabilities[4] such that $\sum_i p_i = 1$ since we are minimizing an expectation of the hinge loss.

$$\min_{w,b} \quad \sum_i p_i[-Y_i(w^T X_i + b) + 1]^+ - zg(w) \tag{4}$$

3.2 bPOE minimization and the bSVM

For our framework, we approach the problem by minimizing bPOE, yielding a regularized formulation with an interpretable parameter and objective function. First, we consider the random loss $\frac{-Y(w^T X+b)}{g(w)}$. We include a non-negative, convex, positive homogenous (PH) term in the denominator to penalize classifiers with large complexity. We can also interpret this as a normalized loss distribution, normalized via $g(w)$. We would still like to minimize POE, since

$$P(\frac{-Y(w^T X + b)}{g(w)} > 0) = P(-Y(w^T X + b) > 0) = P(sign\{w^T X + b\} \neq Y) \ .$$

This is difficult to optimize and also does not take into account the magnitude of the errors, so we instead minimize bPOE, the minimal quasiconvex upper bound of POE. We also consider non-zero thresholds $z \in \mathbb{R}$ so that we can perform some type of model selection. Thus, we find the classifier (w, b) that minimizes bPOE of the normalized loss distribution $\frac{-Y(w^T X+b)}{g(w)}$ at threshold z by solving the following problem, which we call the bSVM.

$$\min_{w,b} \quad \bar{p}_z(\frac{-Y(w^T X + b)}{g(w)}). \tag{5}$$

[2]In the case of using a norm for regularization, the geometric concept of a margin is used to interpret the regularization.

[3]A function $f : \mathbb{R}^n \to \mathbb{R}$ is positive homogenous if $af(w, X) = f(aw, X)$ for any $a \geq 0$.

[4]These are typically assumed to be $\frac{1}{N}$, i.e. equally probable realizations.

Table 1: List of SVM's and how their structure corresponds to (7). Note that all $g(w)$ are convex, PH, and non-negative and thus fit the requirements of Proposition 2. See Appendix for brief description of these SVM's, related references, and explanation of notation.

SVM Formulation	$f(w,X)$	$g(w)$	p_i
C-SVM	$-Y(w^TX+b)$	$\|w\|$	$\frac{1}{N}$
2C-SVM	$-Y(w^TX+b)$	$\|w\|$	$\frac{P(Y=Y_i)}{2}$
Cost Sensitive SVM	$-Y(w^TX+b)$	$\|w\|$	$p_i \mid \sum_i p_i = 1$
ElasticNet-SVM	$-Y(w^TX+b)$	$\lambda\|w\|_1 + (1-\lambda)\|w\|_2$	$\frac{1}{N}$
Laplacian-SVM	$-Y(w^TX+b)$	$\lambda\|w\| + (1-\lambda)\|w^TDM\|_2$	$\frac{1}{N}$
Sparsity Preserving SVM	$-Y(w^TX+b)$	$\lambda\|w\| + (1-\lambda)\|w^TX(I-S)\|_2$	$\frac{1}{N}$
RankSVM	$-w^T(X^+ - X^-)$	$\|w\|$	$\frac{1}{N}$

This can be posed equivalently as the following[5]:

$$\min_{w,b} \quad \sum_i p_i[-Y(w^TX+b) - zg(w) + 1]^+ \tag{6}$$

3.3 SVM's are bPOE minimization with $z \leq 0$

Here, we show via Proposition 2 that (4) and the bSVM (6) are equivalent over the parameter range $z \leq 0$. This result implies that a wide variety of SVM's are equivalent to bPOE minimization. Table 1 lists some example SVM's and shows how they fit into the form of (7) and (8). Note that Proposition 2 proves a more general case, of which (4) and the bSVM are special cases, where we replace $-Y(w^TX+b)$ with a more general convex, PH function $f(w,X)$. Thus, instead of (4), we have (7).

$$\min_{w} \quad \sum_i p_i[f(w,X) + 1]^+ - zg(w) \tag{7}$$

Furthermore, instead of (5) and (6), we have (8) and its equivalent representation (9).

$$\min_{w} \quad \bar{p}_z(\frac{f(w,X)}{g(w)}) \tag{8}$$

$$\min_{w} \quad \sum_i p_i[f(w,X_i) - zg(w) + 1]^+ \tag{9}$$

Proposition 2. *Assume that $f(w,X)$ and $g(w)$ are real valued, convex functions that are positive*

[5]To see this equivalence, rewrite (8) with the denominator as a constraint $g(w) = 1$, apply (3), and make the change of variable $w_{new} = \frac{w}{z-\gamma}$. For details, see a similar proof from Section 4.3 of Norton et al. (2015).

homogenous w.r.t. $w \in \mathbb{R}^n$ *and that* $p_i \in [0,1]$ *such that* $\sum_i p_i = 1$. *Furthermore, assume that* $g(w)$ *is non-negative. Then, over the parameter range* $z \leq 0$, *(9) and (7) produce the same set of optimal solutions.*

Proof. We omit the lengthy proof for brevity and refer readers to the proofs of theorems in Section 6 of Norton et al. (2015). The proof structure is identical except that Norton et al. (2015) consider the special case of $f(w, X) = -Y(w^T X + b)$ and $g(w) = \|w\|$. In principal, the proof works by analyzing the KKT systems of (9) and (7) and finding the appropriate choice of parameters so that the KKT systems become equivalent. □

3.4 New Interpretations of the SVM

The bSVM formulation, being simply bPOE minimization, is much easier to interpret than (4). Specifically, have the following property which shows that the optimal objective value, free parameter, and classification 'margin' have exact interpretations as statistical quantities relating to the optimal loss distribution. This follows directly from Property 1.

Property 3. *Assume for* $z \in \mathbb{R}$, *that* $1 - \alpha^* = \min_{w,b} \sum_i p_i[-Y_i(w^T X_i + b) - zg(w) + 1]^+$.
Then for the normalized loss, $F := \left(\frac{-Y(w^{*T} X + b^*)}{g(w^*)}\right)$, *at the optimal point* (w^*, b^*)*:*

$$\bar{p}_z(F) = 1 - \alpha^*, \quad \bar{q}_{\alpha^*}(F) = z, \quad q_{\alpha^*}(F) = z - \frac{1}{g(w^*)} \quad .$$

First, notice that the optimal objective value equals a probability level. Second, notice that the free parameter is bPOE threshold, or equivalently the superquantile at probability level α^*. Third, we have a new interpretation for what is typically called the 'margin' when $g(w) = \|w\|$. Specifically, we have that the margin is the difference between the superquantile and quantile of the optimal loss distribution,

$$\frac{1}{g(w^*)} = z - q_{\alpha^*}(F) \ .$$

Additionally, note that in formulating the bSVM we do not explicitly utilize the hinge loss function as an approximation or upper bound of the 0-1 loss. The hinge loss naturally arises from minimizing bPOE, e.g. see (6) for $z = 0$. This is an interesting observation, as it is suggested in Rosasco et al.

(2004) that the hinge loss is the best convex approximation to the 0-1 loss. The fact that the hinge loss is a byproduct of minimizing bPOE, the minimal quasiconvex upper bound of POE, shows that it can indeed be viewed as an optimal convex approximation to the 0-1 loss.

4 Robust bSVM and SVM's with Uncertainty

In this section, we consider a robust form of bPOE minimization. Specifically, we consider a simplified case of the following general problem where we minimize bPOE at threshold $z = 0$ where our random vector X is subjected to best-case optimistic uncertainty $\delta^O \in \mathcal{C}^O$ and worst-case pessimistic uncertainty $\delta^P \in \mathcal{C}^P$ with $\mathcal{C}^O, \mathcal{C}^P$ being sets of random vectors.

$$\begin{aligned} \min_{w,\delta^O} \quad \max_{\delta^P} \quad & \bar{p}_0\left(f(w, X + \delta^O + \delta^P)\right) \\ s.t. \quad & \delta^O \in \mathcal{C}^O \\ & \delta^P \in \mathcal{C}^P \ . \end{aligned} \tag{10}$$

With some simplifying assumptions, we can present this problem in the traditional robust optimization framework, optimizing over fixed uncertainty sets. We call this new formulation the Robust bSVM (RObSVM). We show that this formulation is relevant for the following reasons. First, the RObSVM gives us a formulation with many degrees of flexibility that allow us to combine multiple strategies for robust classifier design in an intuitive way. Second, we show that the bSVM is a special case of the RObSVM, allowing us to pose equivalent RO formulations for any problem posed as (4) or (6). Third, we see that many Robust SVM's are special cases of the RObSVM. Fourth, this provides a unique perspective for the convex and non-convex case of the bSVM and RObSVM, specifically showing that the convex case can be viewed as taking a pessimistic view of uncertainty while the non-convex case can be interpreted as taking an optimistic view of uncertainty.

4.1 Robust bSVM

The RObSVM is formulated by simplifying (10) with the following assumptions which allow us to consider, individually, uncertainty about each observation X_i of X as well as formulate this uncer-

tainty as convex sets. Proposition 3 will further clarify the reasoning for these assumptions.

First, we let $f(w, X + \delta^O + \delta^P) := -Y(w^T(X + \delta^O + \delta^P) + b)$. Second, we assume we have N observations $(X_1, Y_1), ..., (X_N, Y_N)$. Third, we assume that every $\delta^O \in \mathcal{C}^O$ and $\delta^P \in \mathcal{C}^P$ are discrete random vectors with N outcomes $\delta_1^O, ..., \delta_N^O$ and $\delta_1^P, ..., \delta_N^P$ where $P(\delta = \delta_i | X = X_i) = 1$. Fourth, we confine each outcome to a closed convex set such that $\delta_i^O \in C(g_i^O, z_i^O)$ and $\delta_i^P \in C(g_i^P, z_i^P)$ for every $i = 1, ..., N$ where $C(g_i^O, z_i^O), C(g_i^P, z_i^P)$ denote closed convex sets. To form these closed convex sets, we first specify convex PH functions $g_i^P : \mathbb{R}^n \to \mathbb{R}$ and $g_i^O : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, ..., N$ and non-negative parameters $z_i^P, z_i^O \geq 0$. Then, for any $z \geq 0$, $g : \mathbb{R}^n \to \mathbb{R}$, we define the closed convex set $C(g, z) \subseteq \mathbb{R}^n$ to be the closed convex set which has support function zg. Recall that a closed convex set C has support function g if $g(w) = \max_{\delta \in C} w^T \delta$ for any $w \in \mathbb{R}^n$. Thus, following from Corollary 13.2.1 in Rockafellar (2015), we know that given any convex PH function gz, it is the support function for the closed convex set,

$$C(g, z) = \{\delta \in \mathbb{R}^n | \forall \hat{\delta} \in \mathbb{R}^n, \delta^T \hat{\delta} \leq zg(\hat{\delta})\} .$$

Making these assumptions, (10) becomes the following formulation, which we call the RObSVM.

$$\begin{aligned} \min_{w, \delta_i^O} \quad \max_{\delta_i^P} \quad & \bar{p}_0 \left(-Y(w^T(X + \delta^O + \delta^P) + b)\right) \\ s.t. \quad & \delta_i^O \in C(g_i^O, z_i^O) \, , i = 1, ..., N \\ & \delta_i^P \in C(g_i^P, z_i^P) \, , i = 1, ..., N. \end{aligned} \tag{11}$$

Intuitively, we individually design each of the $2N$ uncertainty sets by making two choices. First, we choose the function g which defines the *shape* of the convex set $C(g, z)$. Second, we vary the *size* of $C(g, z)$ by choosing the scaling parameter $z \geq 0$. The primary reason for defining the sets $C(g, z)$ in terms of a support function is that it allows us to reformulate the RObSVM in the following way.

Proposition 3. *Assume we have realizations* $(X_1, Y_1), ..., (X_N, Y_N)$ *and probabilities* $p_i \in [0, 1]$, $\sum_i p_i = 1$. *The RObSVM, (11), can be reformulated as (12) without changing the optimal objective*

value or optimal variables w, b.

$$\min_{w,b} \quad \sum_i p_i[-Y_i(w^T X_i + b) + z_i^P g_i^P(w) - z_i^O g_i^O(w) + 1]^+ \tag{12}$$

Proof. Applying bPOE formula (3) to (11), bringing the minimization and maximization w.r.t. δ^O and δ^P inside the summation, and simplifying yields,

$$\min_{w,b} \quad \sum_i p_i[-Y_i(w^T X_i + b) + \min_{\delta_i^O \in C(g_i^O, z_i^O)} (-Y_i w^T \delta_i^O) + \max_{\delta_i^P \in C(g_i^P, z_i^P)} (-Y_i w^T \delta_i^P) + 1]^+ . \tag{13}$$

By definition, zg is the support function of $C(g, z)$, meaning that $\max_{\delta \in C(g,z)} w^T \delta = zg(w)$. Therefore,

$$\min_{\delta_i^O \in C(g_i^O, z_i^O)} -Y w^T \delta_i^O = \min_{\delta_i^O \in C(g_i^O, z_i^O)} (w^T \delta_i^O) = - \max_{\delta_i^O \in C(g_i^O, z_i^O)} -w^T \delta_i^O = - \max_{\delta_i^O \in C(g_i^O, z_i^O)} w^T \delta_i^O = -z_i^O g(w)$$

and $\max_{\delta_i^P \in C(g_i^P, z_i^P)} -Y_i w^T \delta_i^P = z_i^P g_i^P(w)$. Using this to simplify (13) yields (12). □

4.2 The bSVM as a special case

Using Proposition 3, it is then easy to see that the bSVM is a special case of the RObSVM where each disturbance δ_i is confined to the same uncertainty set. Specifically, assume the bSVM is formulated with g and $z \leq 0$, and is thus the convex case. Then, the bSVM is equivalent to the RObSVM with only pessimistic disturbances, where $\delta_i^P \in C(g, -z)$ for every $i = 1, ..., N$ and $\delta_i^O = 0$ for every $i = 1, ..., N$. On the other hand, assume that the bSVM is formulated with g and $z > 0$, and is thus the non-convex case. Then, the bSVM is equivalent to the RObSVM with only optimistic disturbances, where $\delta_i^O \in C(g, z)$ for every $i = 1, ..., N$ and $\delta_i^P = 0$ for every $i = 1, ..., N$.

This result is enlightening for any SVM fitting the form of (4). Proposition 2 showed that these SVM's are equivalent to the bSVM with $z \leq 0$. Thus, using the RObSVM we can formulate robust representations of these SVM's by forming the set $C(g, z)$ which has zg as its support function. A connection of this kind between robustness and regularization has been explored in Xu et al. (2009), but this is a more extensive connection between the choice of regularizer and the implied

Table 2: List of SVM's and the uncertainty set $C(g_i, z_i)$ implied by choice of $g_i(w)$ and z_i. See Appendix for brief description of these SVM's, related references, and explanation of notation.

SVM Formulation	$f(w, X)$	$g_i(w)$	$C(g_i, z_i)$
C-SVM & 2C-SVM & Cost-SVM	$-Y(w^T X + b)$	$\|w\|$	$\{\delta \mid \|\delta\|^* \leq \|z\|\}$
Robust Cost-SVM	$-Y(w^T X + b)$	$\|w\|$	$\{\delta \mid \|\delta\|^* \leq \|z_i\|\}$
Robust ElasticNet-SVM	$-Y(w^T X + b)$	$\lambda\|w\|_1 + (1-\lambda)\|w\|_2$	$\{\delta \mid \lambda\|\delta\|_\infty + (1-\lambda)\|\delta\|_2 \leq \|z_i\|\}$
Laplacian-SVM	$-Y(w^T X + b)$	$\lambda\|w\| + (1-\lambda)\|w^T DM\|_2$	$\{\delta \mid \delta^T\hat{\delta} \leq \|z\|(\lambda\|\ \hat{\delta}\| + (1-\lambda)\|\hat{\delta}^T DM\|_2)\ \ ,\forall\ \hat{\delta} \in \mathbb{R}^n\}$
Sparsity Preserving SVM	$-Y(w^T X + b)$	$\lambda\|w\| + (1-\lambda)\|w^T X(I-S)\|_2$	$\{\delta \mid \delta^T\hat{\delta} \leq \|z\|(\lambda\|\ \hat{\delta}\| + (1-\lambda)\|\hat{\delta}^T X(I-S)\|_2)\ \ ,\forall\ \hat{\delta} \in \mathbb{R}^n\}$
RankSVM	$-w^T(X^+ - X^-)$	$\|w\|$	$\{\delta \mid \|\delta\|^* \leq \|z\|\}$
Ellipsoid Uncertainty SVM	$-Y(w^T X + b)$	$\lambda\|w\| + (1-\lambda)\|\Sigma^{\frac{1}{2}} w\|_2$	$\{\delta \mid \delta^T\Sigma_i^{-1}\delta \leq z_i\}$
Interval Uncertainty SVM	$-Y(w^T X + b)$	$\lambda\|w\| + (1-\lambda)\|S_i w\|_1$	$\{\delta \mid \delta^T\hat{\delta} \leq \|z_i\|(\lambda\|\ \hat{\delta}\| + (1-\lambda)\|S_i\hat{\delta}\|_1)\ \ ,\forall\ \hat{\delta} \in \mathbb{R}^n\}$

uncertainty set. In Table 2, we list examples of SVM's that are special cases of the RObSVM and show the types of uncertainty sets that characterize their robust equivalents. Additionally, note that some examples in Table 2 are special cases of the RObSVM but are not special cases of the bSVM. They utilize different uncertainty sets for each disturbance δ_i, $i = 1, ..., N$ by choosing different g_i and z_i. For example, the Robust Cost-SVM from Katsumata and Takeda (2015) selects z_i individually to reflect cost considerations and the Interval Uncertainty considered by El Ghaoui et al. (2003) selects g_i individually to consider different uncertainty intervals for every sample.

5 Non-Convex bSVM and RObSVM, Extending SVM's with Optimism

In previous sections, we primarily considered the convex case of the RObSVM and bSVM in relation to equivalent SVM's. These formulations, though, also have non-convex cases for particular choice of parameter z which we briefly discuss in this section. Specifically, for the bSVM it is convex when $z \leq 0$ and non-convex when $z > 0$. For the RObSVM, it is convex when only pessimistic disturbances are added to the data and non-convex when optimistic disturbances are additionally considered. Thus, for SVM's fitting into the form of (7), (4), or fitting the convex case of the RObSVM, we can view the bSVM and RObSVM as *extended* formulations. Over the non-convex case, the bSVM extends the allowable range of parameter and can achieve a larger variety of optimal hyperplanes. The RObSVM, over the non-convex case, considers optimistic uncertainty, extending the set of achievable hyperplanes.

Special cases of the non-convex bSVM and RObSVM do exist in the SVM literature, although they are far less common due to the difficulty in optimizing the non-convex objective function. Non-convex extensions of SVM's have been developed for the C-SVM. Furthering work done by

Pérez-Cruz et al. (2003) which developed the Extended ν-SVM (Eν-SVM), a non-convex extension of the ν-SVM proposed in Schölkopf et al. (2000), the Extended C-SVM (EC-SVM) was proposed by Norton et al. (2015). The formulation is exactly the bSVM with $g(w) = \|w\|$ and $p_i = \frac{1}{N}$. Furthermore, in the robust context, Zhang (2005) proposed the Total Support Vector Classifier (TSVC), a non-convex robust SVM utilizing norm based uncertainty to deal with noise. This formulation is exactly the RObSVM with only norm constrained optimistic disturbances $\|\delta_i^O\| \leq z_i^O$ for every $i = 1, ..., N$. Although difficult to optimize, evidence was presented that these formulations performed better than their convex counterparts on some data sets. In particular, Zhang (2005) showed that the TSVC was robust to noisy input data. In the context of our proposed scheme and interpretation, the performance of the TSVC makes sense. If data are corrupted with noise, perhaps an optimistic view of the data would produce better classification.

Future work: Sparsity inducing regularizers and non-convex extensions Recently, non-convex regularization schemes have shown great success in inducing sparsity, see e.g. Tono et al. (2017), Gotoh et al. (2015), Yin et al. (2015). Therefore, it is interesting to note that some of these regularization schemes can be represented using our framework as the RObSVM formulated with a combination of optimistic and pessimistic uncertainty, yielding a non-convex regularizer $z_i^P g_i^P(w) - z_i^O g_i^O(w)$. Thus, it would be interesting to fully explore this new interpretation for non-convex regularization and to propose new non-convex regularization schemes.

Given the numerical success of extended, non-convex formulations presented in Pérez-Cruz et al. (2003), Zhang (2005), we feel that it would also be beneficial to explore other, more flexible, non-convex formulations. For example, our formulation suggests that the Laplacian-SVM has a non-convex extension. Numerical studies should be conducted to determine the use and effectiveness of such extended formulations. In addition, the recent success of non-convex optimization techniques, such as DCA Dinh and Le Thi (2014), present a path toward practical solution methods.

6 Conclusion

In this paper, we approached the binary classification problem by minimizing bPOE, the minimal quasiconvex approximation to POE. We showed that a variety of regularized and robust SVM's are special cases of this approach. This new perspective on SVM's is beneficial, providing a statistical interpretation for its optimal objective value, its free parameter, and the margin induced by the choice of regularizer. Additionally, we are able to provide a strong connection between regularization and robustness that extends the current understanding in the literature. We show that regularized SVM's can be cast as robust SVM's where data uncertainty is characterized by a convex uncertainty set which has support function equal to the regularization function scaled by the value of the free parameter z. Additionally, we show that the bPOE minimization problems have both convex and non-convex cases, extending the set of optimal hyperplanes achievable by related convex SVM formulations. Furthermore, we find that the robust perspective provides an intuitive view of the convex and non-convex case as related to optimistic or pessimistic views of the data set. Overall, we find that bPOE minimization provides a unifying view of many SVM variants, including regularized, robust, convex, and non-convex formulations.

A Overview of SVM formulations in Table 1 and 2

Within the Table 1 and 2, let $\|\cdot\|$ denote a general norm, let $\|\cdot\|^*$ denote its dual norm, let $\|\cdot\|_p$ denote the L_p norm, and let $\lambda \in [0,1]$ be used to denote a convex combination. Notation that is specific to certain SVM's can be found in the following overview of formulations.

- C-SVM: Original formulation from Cortes and Vapnik (1995).
- 2C-SVM: SVM for unbalanced classes from Osuna et al. (1997).
- ElasticNet-SVM: Doubly regularized SVM from Wang et al. (2006), with $\lambda \in [0,1]$.
- Laplacian-SVM: Semi-Supervised Manifold regularization from Belkin et al. (2006). We have a set of unlabeled, U, and labeled, X, data $D = [X, U]$. We form and decompose the Laplacian $L = MM^T$ (see e.g. Von Luxburg (2007) for a tutorial on the Laplacian). We form the regularizer $g(w) = \lambda\|w\| + (1-\lambda)\|w^T DM\|$ where $\lambda \in [0,1]$.
- Sparsity Preserving-SVM: Manifold Regularization scheme from Shen et al. (2015). Similar to Laplacian-SVM, but with I denoting the identity matrix and S a sparse representation of the data set X found via sparse coding.

- Rank SVM: AUC maximizing SVM from Herbrich et al. (1999), Rakotomamonjy (2004), Brefeld and Scheffer (2005). X^+, X^- denote random vectors with positive and negative label respectively.
- Robust Cost-SVM: Cost sensitive formulation from Katsumata and Takeda (2015) putting larger uncertainty around minority class data vectors.
- Robust ElasticNet-SVM: Robust variant of ElasticNet from Katsumata and Takeda (2015).
- Ellipsoid Uncertainty SVM: SVM considering ellipsoid uncertainty with ellipsoid shape matrix Σ^{-1} from Trafalis and Gilbert (2007), Bhattacharyya et al. (2005).
- Interval Uncertainty SVM: SVM considering hyper-rectangular uncertainty sets from El Ghaoui et al. (2003).

References

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.

Chiranjib Bhattacharyya, KS Pannagadatta, and Alexander J Smola. A second order cone programming formulation for classifying missing data. In *Neural Information Processing Systems (NIPS)*, pages 153–160, 2005.

U. Brefeld and T. Scheffer. AUC maximizing support vector learning. *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, 2005.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Tao Pham Dinh and Hoai An Le Thi. Recent advances in dc programming and dca. In *Transactions on Computational Intelligence XIII*, pages 1–37. Springer, 2014.

Laurent El Ghaoui, Gert René Georges Lanckriet, Georges Natsoulis, et al. Robust classification with interval data, 2003.

Jun-ya Gotoh, Akiko Takeda, and Katsuya Tono. Dc formulations and algorithms for sparse optimization problems. *Preprint, METR*, 27, 2015.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999.

Shuichi Katsumata and Akiko Takeda. Robust cost sensitive support vector machine. In *AISTATS*, 2015.

A. Mafusalov and S. Uryasev. Buffered probability of exceedance: Mathematical properties and optimization algorithms. *Research Report 2014-1, ISE Dept., University of Florida*, 2015.

M. Norton and S. Uryasev. Maximization of auc and buffered auc in binary classification. *Research Report 2014-2, ISE Dept., University of Florida*, 2014.

M. Norton, A. Mafusalov, and S. Uryasev. Soft margin support vector classification as buffered probability minimization. *Research Report 2015-2, ISE Dept., University of Florida*, 2015.

Edgar Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training and applications. 1997.

Fernando Pérez-Cruz, Jason Weston, DJL Herrmann, and B Scholkopf. Extension of the nu-svm range for classification. *NATO SCIENCE SERIES SUB SERIES III COMPUTER AND SYSTEMS SCIENCES*, 190:179–196, 2003.

Alain Rakotomamonjy. Optimizing area under roc curve with svms. In *ROCAI*, pages 71–80. Citeseer, 2004.

Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

R.T. Rockafellar and J.O. Royset. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety, Vol. 95, 499-510*, 2010.

R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk, Vol. 2, No. 3, 2000, 21-41*, 2000.

Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

Bin Shen, Bao-Di Liu, Qifan Wang, Yi Fang, and Jan P Allebach. Sp-svm: Large margin classifier for data on multiple manifolds. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Katsuya Tono, Akiko Takeda, and Jun-ya Gotoh. Efficient dc algorithm for constrained sparse optimization. *arXiv preprint arXiv:1701.08498*, 2017.

Theodore B Trafalis and Robin C Gilbert. Robust support vector machines for classification and computational issues. *Optimisation Methods and Software*, 22(1):187–198, 2007.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Li Wang, Ji Zhu, and Hui Zou. The doubly regularized support vector machine. *Statistica Sinica*, pages 589–615, 2006.

Huan Xu, Constantine Caramanis, Shie Mannor, and Sungho Yun. Risk sensitive robust support vector machines. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 4655–4661. IEEE, 2009.

Penghang Yin, Yifei Lou, Qi He, and Jack Xin. Minimization of 1-2 for compressed sensing. *SIAM Journal on Scientific Computing*, 37(1):A536–A563, 2015.

Jinbo Bi Tong Zhang. Support vector classification with input data uncertainty. *Advances in neural information processing systems*, 17:161, 2005.