Soft Margin Support Vector Classification as Buffered Probability Minimization

Matthew Norton, Alexander Mafusalov, Stan Uryasev

July 21, 2016

Matthew Norton, *mdnorton@ufl.edu*, PhD Student, Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, 303 Weil Hall, University of Florida, Gainesville, FL 32611.
Alexander Mafusalov, *mafusalov@ufl.edu*, PhD Candidate, Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, 303 Weil Hall, University of Florida, Gainesville, FL 32611.
Stan Uryasev, *uryasev@ufl.edu*, George & Rolande Willis Endowed Professor, Director of Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, 303 Weil Hall, University of Florida, Gainesville, FL 32611.

RESEARCH REPORT 2015-2
Risk Management and Financial Engineering Lab
Department of Industrial and Systems Engineering
303 Weil Hall, University of Florida, Gainesville, FL 32611.
E-mail: *mdnorton@ufl.edu, mafusalov@ufl.edu, uryasev@ufl.edu*

First Draft: August 2015, This Draft: January 2016

Abstract

In this paper, we show that the popular C-SVM, soft-margin support vector classifier is equivalent to minimization of Buffered Probability of Exceedance (bPOE) by introducing a new SVM formulation, called the EC-SVM, which is derived as a bPOE minimization problem. Since it is derived from a simple bPOE minimization problem, the EC-SVM is simple to interpret with a meaningful free parameter, optimal objective value, and probabilistic derivation. We connect the EC-SVM to existing SVM formulations. We first show that the C-SVM, formulated with any regularization norm, produces the same set of solutions as the EC-SVM over the same parameter range. Additionally, we show that the Eν-SVM, formulated with any regularization norm, produces the same set of solutions as the EC-SVM over their entire parameter range. These equivalences, coupled with the interpretability of the EC-SVM, allow us to gain surprising new insights into the C-SVM and fully connect soft margin support

vector classification with superquantile and bPOE concepts. Additionally, we provide general dual formulations for the EC-SVM and C-SVM allowing for a brief discussion of application of the kernel trick to the EC-SVM.

Keywords: Support Vector Machines, Buffered Probability of Exceedance, Conditional Value-at-Risk, Binary Classification, Machine Learning

1 Introduction

In the machine learning community, the Soft Margin Support Vector Machine (C-SVM) has proven to be an extremely popular tool for classification, spawning generalizations for regression, robust optimization, and a host of other applications. With its connection to statistical learning theory, intuitive geometric interpretation, and efficient extensions to non-linear classification, the C-SVM has proven to be a flexible tool based on sound theory and intuition. Still, there are insights left to be gained with regard to soft margin support vector classification.

One such insight was revealed by [12], where it was shown that the Eν-SVM, an extension of the ν-SVM, is equivalent to superquantile minimization. Superquantiles, popularized in the financial engineering literature under the name Conditional Value-at-Risk (CVaR), were developed by [8] as means for dealing with optimization of quantiles. Utilizing the popular calculation formula for superquantiles, Takeda showed that the Eν-SVM was equivalent to superquantile minimization, with the free parameter of the Eν-SVM being equivalent to the free choice of probability level in superquantile minimization.

In this paper, we provide insights in a similar direction by utilizing the inverse of the superquantile, so called buffered probability of exceedance (bPOE). More specifically, bPOE is a generalization of *buffered Probability of Failure* (bPOF) which was introduced by [9] and further studied in [10]. This generalization, recently studied by [4, 6, 5, 3, 13], has shown a great deal of promise as generating numerically tractable methods for probability minimization.

Utilizing the bPOE concept, we introduce a new SVM formulation called the Extended Soft Margin Support Vector Machine (EC-SVM). Being derived as a bPOE minimization problem, the EC-SVM is simple to interpret. First, we show that the EC-SVM has a free parameter interpretable as a specific statistical quantity relating to the optimal loss distribution. Second, we show that the value of the optimal objective function (divided by sample size) is equal to a probability level. Lastly, we show that the EC-SVM can be interpreted as having a hard-margin criterion. Additionally, with the EC-SVM formulated with any general norm, we show that the choice of norm implies a distance metric which defines the hard-margin criterion.

After introducing the EC-SVM, we then connect it to existing SVM formulations. In our main result, we show that the C-SVM and EC-SVM, when formulated with any general norm and non-negative parameter values, produce the same set of optimal hyperplanes.

This result implies that the original soft-margin SVM formulation, derived in great part from geometric intuition, is equivalent to minimization of bPOE, a probabilistic concept. This result also implies that the interpretation of the EC-SVM's parameter, optimal objective, and hard-margin criterion can be applied to the C-SVM. This includes the surprising result that the optimal objective value of the C-SVM, divided by sample size, equals a probability level.

We also connect the EC-SVM and Eν-SVM, showing that these SVM formulations produce the same set of optimal hyperplanes over their entire parameter range. With bPOE being the inverse of the superquantile, this relationship follows immediately from the derivation of the EC-SVM as a bPOE minimization problem and the results of Takeda. This result also makes it clear that the EC-SVM is an extension of the C-SVM in the same way that the Eν-SVM is an extension of the ν-SVM.

Additionally, we provide the dual formulation of the EC-SVM and C-SVM formulated with any general regularization norm. This allows us to briefly discuss application of the kernel trick to the EC-SVM, a popular operation considered for the C-SVM in the case of the L_2 regularization norm.

This paper is organized as follows. Note that in order to make this paper as self-contained as possible, we include a significant amount of review in the first three sections. Section 2 reviews some existing SVM formulations relevant to our discussion, specifically the C-SVM, ν-SVM, and Eν-SVM. Section 3 briefly reviews the concept of a superquantile and the results of Takeda, which show that the Eν-SVM is equivalent to superquantile minimization. Section 4 reviews the bPOE concept, which is critical to our contribution. Additionally, we present a new formulation for minimizing bPOE in the presence of Positive Homogenous (PH) random functions. The necessity of this new formulation is discussed in more detail in Appendix A. Section 5 introduces the EC-SVM as a bPOE minimization problem, and discusses the properties of the EC-SVM and its interpretation as a hard-margin optimization problem. Section 6 connects the C-SVM and EC-SVM. Section 7 connects the EC-SVM and Eν-SVM, and presents the results of this and previous papers in a cohesive framework connecting soft margin support vector classification and superquantile concepts. Section 8 presents dual formulations on the C-SVM and EC-SVM formulated with any general norm, discusses application of the kernel trick, and shows that the optimal objective of the C-SVM, divided by sample size, equals a probability level.

2 C-SVM, ν-SVM, Eν-SVM

In this section, we review three existing SVM formulations; C-SVM, ν-SVM, and Eν-SVM. We begin with a review of the C-SVM and ν-SVM, reviewing the fact that they share the same optimal solution sets. We then review the interpretation of the ν-SVM parameter, its limitations, and the Eν-SVM formulation which serves to resolve these limitations.

2.1 The C-SVM

Consider the task of binary classification where we have a set of N feature vectors $X_i \in \mathbb{R}^n$ and associated class labels $y_i \in \{-1, +1\}$ and we need to choose a hyperplane $w \in \mathbb{R}^n$ with intercept $b \in \mathbb{R}$ to properly classify feature vectors via the linear decision function $d(w, b, X_i) = sign(w^T X_i + b)$.

One of the most successful algorithms for accomplishing this task is the soft-margin SVM, also referred to as the C-SVM. The C-SVM is formulated as (1), where $C \geq 0 \in \mathbb{R}$ is chosen as a fixed tradeoff parameter and the norm is typically the L_2, $\|\cdot\|_2$, or L_1, $\|\cdot\|_1$, norm. Below, we present it with the general norm, $\|\cdot\|$.

$$\begin{aligned} \min_{w,b,\xi} \quad & C\|w\| + \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i + b) + 1, \quad \forall i \in \{1, ..., N\}, \\ & \xi \geq 0. \end{aligned} \tag{1}$$

2.2 The ν-SVM

After introduction of the C-SVM, the ν-SVM was introduced as an equivalent formulation with more intuitive parameter choices. C-SVM and ν-SVM, with the L_2 norm, are equivalent in that they provide the same set of optimal solutions over the space of all possible parameter choices (see [2]). These algorithms are different, though, in the meaning of the value of the free parameter. For the C-SVM, there was no direct interpretation for the meaning of the C-parameter other than as a trade-off between margin size and classification errors. The ν-SVM, on the other hand, provided a more interpretable parameter.

The ν-SVM is traditionally formulated as (2) with the L_2 norm, where $\nu \in [0, 1]$ instead of $C \in [0, +\infty)$ is chosen as a fixed tradeoff parameter.

$$\begin{aligned} \min_{w,b,\rho,\xi} \quad & \frac{1}{2}\|w\|_2^2 - \nu\rho + \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i + b) + \rho, \quad \forall i \in \{1, ..., N\}, \\ & \xi \geq 0. \end{aligned} \tag{2}$$

As already mentioned, the ν-SVM advantageously gives us a free parameter, $\nu \in [0, 1]$ with implied meaning. The meaning of ν is shown in [11] by proving a variant of Property 1.

Property 1. *Assume that there exists a feasible solution (w, b, ρ, ξ) for (2) for parameter choice $\nu \in [0, 1]$. Then the following bounds apply:*

- *The choice of ν acts as an upper bound on the fraction of errors in the margin:*

$$1 - \alpha < \nu, \quad where \quad 1 - \alpha = \frac{1}{N}\left|\{i : y_i(w^T X_i + b) < \rho\}\right|.$$

- *The choice of ν acts as a lower bound on the fraction of support vectors (SV's), support vectors being errors that lie in the margin or on the margin boundary:*

$$\%SV's > \nu, \quad where \quad \%SV's = \frac{1}{N}\left|\{i : y_i(w^T X_i + b) \leq \rho\}\right|.$$

2.3 The Eν-SVM

Given the natural interpretation for the meaning of the ν-parameter, it would seem normal to assume that all the values of $\nu \in [0, 1]$ will yield non-trivial, feasible solutions satisfying the bounds stated in Property 1. This, though, is not the case. In [2], it was shown that the ν-parameter has a limited range. Specifically, a variant of Property 2 is proved in [2].

Property 2.
- *There exists a minimum and maximum value such that $v \in (\nu_{\min}, v_{\max}]$ yields feasible (2) with non-trivial solutions, $\nu \leq \nu_{\min}$ yields (2) with trivial optimal solution $w = b = 0$, and $\nu > \nu_{\max}$ yields infeasible (2).*

- *Furthermore, this limitation in parameter range applies to the C-SVM as well. Specifically, there exists a correspondence in allowable range such that $\nu \to \nu_{\min}$ corresponds to $C \to \infty$ and $\nu \to \nu_{\max}$ corresponds to $C \to 0$.*

To solve this issue, extending the valid range of the ν-parameter to be the entire $[0, 1]$ interval, [7] developed an extended ν-SVM formulation called Eν-SVM (3). The Eν-SVM is traditionally formulated as follows with the L_2 norm. We present it with the general norm as follows:

$$\begin{aligned} \min_{w,b,\rho,\xi} \quad & -\nu\rho + \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i + b) + \rho, \quad \forall i \in \{1, ..., N\}, \\ & \xi \geq 0, \\ & \|w\| = 1. \end{aligned} \tag{3}$$

One can view (3) as an **extension** of (2) in that [7, 2] showed that the optimal solution to (2), formulated with L_2 norm and any $\nu_0 \in (\nu_{min}, v_{max}]$, is also an optimal solution to (3), formulated with L_2 norm, for some $\nu_1 \leq \nu_0$. Problem (3), though, can achieve solutions that (2) cannot because of its extended range of the ν-parameter.

3 Superquantiles and the Eν-SVM

In this section, we first give a brief review of the superquantile concept as introduced by [8]. We then review the results of Takeda, showing that the Eν-SVM is equivalent to superquantile minimization.

3.1 Superquantiles and Tail Probabilities

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 [8] 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

$$\bar{q}_\alpha(Z) = E\left[Z|Z > q_\alpha(Z)\right]. \tag{4}$$

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

$$\bar{q}_\alpha(Z) = \min_{\gamma} \quad \gamma + \frac{E[Z-\gamma]^+}{1-\alpha}, \tag{5}$$

where $[\cdot]^+ = \max\{\cdot, 0\}$. For a discretely distributed random variable Z with equally probable realizations $\{Z_1, Z_2, ..., Z_N\}$ we can write this formula as the following Linear Programming problem

$$\begin{aligned} \bar{q}_\alpha(Z) = \min_{\gamma,\xi} \quad & \gamma + \frac{1}{N(1-\alpha)} \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq Z_i - \gamma, \forall i \in \{1, ..., N\}, \\ & \xi \geq 0. \end{aligned} \tag{6}$$

Similar to $q_\alpha(Z)$, the superquantile can be used to assess the tail of the distribution. The superquantile, though, is far easier to handle in optimization contexts. It also has the important property that it considers the magnitude of events within the tail. Therefore, in situations where a distribution may have a heavy tail, the superquantile accounts for magnitudes of low-probability large-loss tail events while the quantile does not account for this information.

3.2 Eν-SVM as superquantile minimization

In [12], the meaning of the ν-parameter was solidified by showing that the Eν-SVM, (3), is equivalent to superquantile minimization. Specifically, Takeda proved a variant of Prop-

erty 3.

Property 3. *Consider optimization problem (3). Let $\alpha = 1 - \nu$ and $\gamma = -\rho$. Also, let $L(w, b, X, y) = -y(w^T X + b)$ be a discretely distributed random variable with equally probable realizations $\{-y_1(w^T X_1 + b), ..., -y_N(w^T X_N + b)\}$. With this notation, (3) can be rewritten as (7), which is equivalent to (8), minimization of the α-superquantile:*

$$\begin{aligned} \min_{\gamma,\xi} \quad & (1-\alpha)\left(\gamma + \frac{1}{N}\sum_{i=1}^{N}\xi_i\right) \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i + b) - \gamma, \quad \forall i \in \{1, ..., N\}, \\ & \xi \geq 0, \\ & \|w\| = 1. \end{aligned} \tag{7}$$

$$\begin{aligned} \min_{w,b} \quad & (1-\alpha)\bar{q}_\alpha\left(-y(w^T X + b)\right) \\ s.t. \quad & \|w\| = 1. \end{aligned} \tag{8}$$

With this, one can see that the Eν-SVM is simply minimization of the value (5) multiplied by $1 - \alpha$ with the real valued discretely distributed random loss $L(w, b, X, y) = -y(w^T X + b)$ in place of the real valued random variable Z.

4 Buffered Probability of Exceedance (bPOE)

In this section, we first review the concept of bPOE. We show how it is simply one minus the inverse of the superquantile and review its surprising calculation formula. We then review how minimization of bPOE integrates quite nicely into optimization frameworks. Finally, we present a slightly altered formulation for minimization of bPOE in the presence of Positive Homogenous (PH) random functions. For the interested reader, we discuss the necessity of this alteration and derive the formulation in Appendix A. We move this discussion to the appendix, as it is slightly distracting from the discussion related to support vector machines.

4.1 bPOE: Inverse of the superquantile

As mentioned in Section 3, when working with optimization of tail probabilities, one frequently works with constraints or objectives involving POE, $p_z(Z) = P(Z > x)$, 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 superqunatile was developed to alleviate difficulties associated with optimization problems involving quantiles. 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.

Specifically, there exist two slightly different variants of bPOE, namely Lower and Upper bPOE. Paper [4] mainly works with so called Lower bPOE while [6] works with so called Upper bPOE. These definitions do not differ dramatically and a discussion of these differences in beyond the scope of this paper. Thus, for the remainder of this paper when we refer to bPOE, we are utilizing Upper bPOE. With this in mind, bPOE is defined in the following way, where $\sup Z$ denotes the essential supremum of random variable Z. In this paper we assume all random variables to be L_1-finite, $Z \in \mathcal{L}^1(\Omega)$, i.e. $E|Z| < \infty$.

Definition 1. *Upper 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 superquantile equals z. Although bPOE seems troublesome to calculate, [6] 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*

$$\bar{p}_z(Z) = \inf_{\gamma < z} \frac{E[Z-\gamma]^+}{z-\gamma} = \begin{cases} \lim\limits_{\gamma \to -\infty} \frac{E[Z-\gamma]^+}{z-\gamma} = 1, & \text{if } z \leq E[Z], \\ \min\limits_{\gamma < z} \frac{E[Z-\gamma]^+}{z-\gamma}, & \text{if } z \in (E[Z], \sup Z), \\ \lim\limits_{\gamma \to z^-} \frac{E[Z-\gamma]^+}{z-\gamma} = P(Z = \sup Z), & \text{if } z = \sup Z, \\ \min\limits_{\gamma < z} \frac{E[Z-\gamma]^+}{z-\gamma} = 0, & \text{if } \sup Z < z. \end{cases} \tag{9}$$

It is also important to note that formula (9) has the following property, proved in [4] and [6]. The Property 4 will become important in later sections when we begin to interpret the EC-SVM.

Property 4. *If $z \in (E[Z], \sup Z)$ and $\min\limits_{\gamma < z} \frac{E[Z-\gamma]^+}{z-\gamma} = \frac{E[Z-\gamma^*]^+}{z-\gamma^*} = 1-\alpha^*$, then:*

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

Thus, using formula (9), bPOE can be efficiently calculated. Additionally, Property 4 shows that we can recover quantile and superquantile information. As we demonstrate in the next section, formula (9) also allows for convenient optimization of bPOE.

4.2 Optimization of bPOE for random PH functions

Paper [6] considered the following optimization setup to demonstrate the ease with which bPOE can be minimized directly. Assume we have a real valued PH random function $f(w, X)$ determined by a vector of control variables $w \in \mathbb{R}^n$ and a random vector X. By definition, a function $f(w, X)$ is PH w.r.t. w if it satisfies the following condition: $af(w, X) = f(aw, X)$ for any $a \geq 0, a \in \mathbb{R}$.

Now, assume that we would like to find the vector of control variables, $w \in \mathbb{R}^n$, that minimizes the probability of $f(w, X)$ exceeding a threshold of $z \in \mathbb{R}$. We would like to solve the following POE optimization problem.

$$\min_{w \in \mathbb{R}^n} \quad p_z(f(w, X)). \tag{10}$$

Here we have a discontinuous and non-convex objective function (assuming a discretely distributed X) that is numerically difficult to minimize. Consider alternatively minimization of bPOE instead of POE at the same threshold z. This is posed as the optimization problem

$$\min_{w \in \mathbb{R}^n} \quad \bar{p}_z(f(w, X)). \tag{11}$$

Given Proposition 1, (11) can be transformed into the following:

$$\min_{w \in \mathbb{R}^n, \gamma < z} \quad \frac{E[f(w, X) - \gamma]^+}{z - \gamma}. \tag{12}$$

Paper [6], though, limits consideration to only a threshold of $z = 0$, in which case (12) reduces to

$$\min_{w \in \mathbb{R}^n} \quad E[f(w, X) + 1]^+. \tag{13}$$

As we discuss in Appendix A, formulation (12) has shortcomings for nonzero thresholds. Specifically, it fails to achieve varying optimal solutions for varying threshold levels. To address these issues, the next section provides an alternative formulation for bPOE minimization with PH functions $f(w, X)$ that allows effective variation of threshold levels.

4.3 An Altered Formulation

With (12) failing to achieve varying optimal solutions as the threshold z varies, we find that adding a constraint on the norm of w remedies this situation (because w can no longer rescale as the threshold changes). Here $\|\cdot\|$ denotes any general norm. This gives us

$$\begin{aligned} \min_{w \in \mathbb{R}^n, \gamma < z} \quad & \frac{E[f(w, X) - \gamma]^+}{z - \gamma} \\ s.t. \quad & \|w\| = 1. \end{aligned} \tag{14}$$

Furthermore, the following Proposition 4 shows that (14) can be simplified, yielding

$$\min_{w \in \mathbb{R}^n} \quad E[f(w,X) - z\|w\| + 1]^+. \tag{15}$$

Proposition 2. *Assume $f(w,X)$ is PH w.r.t. w. If (w^*, γ^*) is optimal to (14) with optimal objective $1-\alpha^*$, then $w = \frac{w^*}{z-\gamma^*}$ is optimal to (15) with optimal objective $1-\alpha^*$.*

Proof. For this, we show that (15) is formed only by making a change of variable in (14). We start with (14). Since $\gamma < z$ is an explicit constraint and thus $z - \gamma > 0$, we bring the denominator into the expectation in the numerator to get

$$\begin{aligned} \min_{w \in \mathbb{R}^n, \gamma} \quad & E\left[f\left(\frac{w}{z-\gamma}, X\right) - \left(\frac{\gamma}{z-\gamma}\right)\right]^+ \\ s.t. \quad & \|w\| = 1. \end{aligned} \tag{16}$$

Now, make the change of variable $w_{new} = \frac{w}{z-\gamma}$. Since we have the explicit constraint $\|w\| = 1$, we have that $\|w_{new}\| = \left\|\frac{w}{z-\gamma}\right\| = \frac{\|w\|}{z-\gamma} = \frac{1}{z-\gamma}$. We can then make the change of variable to get

$$\min_{w_{new} \in \mathbb{R}^n, \gamma} \quad E\left[f(w_{new}, X) - \|w_{new}\|\gamma\right]^+. \tag{17}$$

We can also rearrange $\|w_{new}\| = \frac{1}{z-\gamma}$ to get $\gamma = z - \frac{1}{\|w_{new}\|}$ and thus that $\|w_{new}\|\gamma = \|w_{new}\|\left(z - \frac{1}{\|w_{new}\|}\right) = z\|w_{new}\| - 1$. Plugging this into our formulation, we arrive at

$$\min_{w_{new} \in \mathbb{R}^n, \gamma} \quad E[f(w_{new}, X) - z\|w_{new}\| + 1]^+, \tag{18}$$

where due to our change in variable, we see that if we have optimal solution w^*_{new}, then $\left(w = \frac{w^*_{new}}{\|w_{new}\|}, \gamma = z - \frac{1}{\|w^*_{new}\|}\right)$ is optimal to (14) before the change of variable. □

Thus, we can turn to (15) as our formulation for bPOE minimization of a PH function for varying threshold choices. Notice from the proof of Proposition 4, that if w^* is optimal to (15) then $(w = \frac{w^*}{\|w^*\|}, \gamma = z - \frac{1}{\|w^*\|})$ is optimal to (14). Also, let us call $f\left(\frac{w^*}{\|w^*\|}, X\right)$ the normalized loss distribution at w^*. Given Property 4, if

$$E[f(w^*, X) - z\|w^*\| + 1]^+ = 1 - \alpha^*,$$

we know the following about the normalized loss distribution at the optimal point w^*:

$$\bar{p}_z(F) = 1 - \alpha^*, \quad \bar{q}_{\alpha^*}(F) = z, \quad q_{\alpha^*}(F) = z - \frac{1}{\|w^*\|}, \quad \text{where } F := f\left(\frac{w^*}{\|w^*\|}, X\right).$$

5 Extended C-SVM

In this section, we use formula (15) to introduce the EC-SVM. We approach the classification problem via a natural bPOE minimization problem. Utilizing the interpretability of optimization problem (15), we show that the EC-SVM is also simple to interpret. Specifically, we show that application of Property 4 allows us to interpret the choice of parameter and the value of the optimal objective in interesting, purely statistical ways.

We also point out that the traditional *hinge loss function* naturally occurs when minimizing bPOE, meaning that we do not need to explicitly specify it as a loss function. This can be seen clearly by considering minimization of bPOE at threshold $C = 0$. Additionally, it is interesting to notice that we do not explicitly attempt to regularize via use of norms, as the use of norms naturally arises from (14).

5.1 bPOE minimization with SVM loss

Consider the formula for bPOE minimization (15) with discretely distributed random loss $L(w, b, X, y) = -y(w^T X + b)$ with equally probable realizations $\{-y_1(w^T X_1 + b), ..., -y_N(w^T X_N + b)\}$ and threshold $z \in \mathbb{R}$. If we let $C = -z$ and multiply the objective function by N, this gives us the following optimization problem, which we call the EC-SVM:

$$\begin{aligned} \min_{w,b,\xi} \quad & \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i + b) + C\|w\| + 1, \quad \forall i \in \{1, \ldots, N\}. \\ & \xi \geq 0. \end{aligned} \tag{19}$$

Notice that the norm $\|\cdot\|$, just as in (15), is an arbitrary norm in $\mathbb{R}^n$. Notice also that the EC-SVM is a very natural formulation. It is simply a *buffered* way of minimizing the probability that misclassification errors exceed our threshold $-C$. Put specifically, instead of finding the classifier (w, b) that minimizes $p_{-C}\left(-y(w^T X + b)\right) = P\left(-y(w^T X + b) > -C\right)$, we are minimizing its buffered variant, $\bar{p}_{-C}\left(-y(w^T X + b)\right)$.

5.2 Occurence of Hinge Loss

When discussing SVM's, it is traditional for people to say that the C-SVM minimizes a *hinge loss function* plus a regularization term on the vector w. For the EC-SVM, though, we see that this is not an accurate descriptor, as we do not need to explicitly specify a hinge loss function. We see that it naturally arises when minimizing bPOE. Specifically, minimizing bPOE of the loss function $L(w, b, X, y) = -y(w^T X + b)$ at threshold $C = 0$ yields the following problem, which is exactly minimization of hinge losses:

$$\min_{w,b} \quad \sum_{i=1}^{N} \left[-y_i(w^T X_i + b) + 1\right]^+. \tag{20}$$

One can see this more generally by looking at (13), where we are minimizing bPOE of a PH loss function $f(w, X)$ at threshold zero.

5.3 Interpretability of EC-SVM

Since the EC-SVM is simply bPOE minimization, we can utilize Property 4 to interpret the C-parameter and the value of the optimal objective in an exact way. Specifically, let us put Property 4 in terms of the EC-SVM.

Property 5. *Suppose (19), with any general norm, yields optimal hyperplane (w^*, b^*) and the optimal objective value equal to obj^*. Considering $L(w^*, b^*, X, y) = -y(w^{*T}X + b^*)$ as a discretely distributed random loss, we know the following about the normalized loss distribution.*

- $\bar{p}_{-C}\left(\frac{L(w^*,b^*,X,y)}{\|w^*\|}\right) = 1 - \alpha^* = \frac{obj^*}{N}$,
- $\bar{q}_{\alpha^*}\left(\frac{L(w^*,b^*,X,y)}{\|w^*\|}\right) = -C$,
- $q_{\alpha^*}\left(\frac{L(w^*,b^*,X,y)}{\|w^*\|}\right) = -C - \frac{1}{\|w^*\|}$.

The C-parameter as superquantile threshold choice. For the C-SVM, the non-negative C-parameter is typically discussed as being a tradeoff between errors and margin size. In the broad scheme of Empirical Risk Minimization (ERM), this parameter is discussed as the tradeoff between risk and regularization.

For the EC-SVM, we provide a much more concrete interpretation. With the EC-SVM being exactly bPOE minimization, the C-parameter is a choice of threshold $z = -C$. Specifically, looking at Property 5, we have that

$$\bar{q}_{\alpha^*}\left(\frac{L(w^*, b^*, X, y)}{\|w^*\|}\right) = -C,$$

where $\alpha^* = 1 - \frac{obj^*}{N} = 1 - \bar{p}_{-C}\left(\frac{L(w^*,b^*,X,y)}{\|w^*\|}\right)$.

The optimal objective value as bPOE. The EC-SVM also has the surprising property that the optimal objective value, divided by the number of samples, is a probability level. More specifically, as shown in Property 5, we have that $\frac{obj^*}{N} = \bar{p}_{-C}\left(\frac{L(w^*,b^*,X,y)}{\|w^*\|}\right)$. In words, the optimal objective value divided by the number of samples equals bPOE of the optimal normalized loss distribution at threshold $-C$.

5.4 Norm choice and margin interpretation

In this section, we show that the EC-SVM has a clear interpretation of 'margin', which is dependent on the choice of norm. This interpretation shows, in an exact way, how the C parameter determines the margin of the separating hyperplane.

Looking back to formula (14), using loss function $f(w, X) = -y(w^T X + b)$, and making the change of variable $w \to \frac{w}{C}, b \to \frac{b}{C}$ we are able to formulate the following equivalent problem:

$$\begin{aligned} \min_{\gamma<-C,w,b} \quad & \sum_{i=1}^{n} \left[\left(\frac{1}{-C-\gamma} \right) \left(\frac{-y_i(w^T X_i + b) + 1}{\|w\|} \right) + 1 \right]^+ \\ s.t. \quad & \frac{1}{\|w\|} = C. \end{aligned} \tag{21}$$

From this formulation, we can form an interpretation of the EC-SVM within the context of selecting an optimal hyperplane under a 'hard margin' criterion. To make this interpretation clear, we can start by analyzing (21) formulated with the L_2 norm, traditional to the C-SVM and discussions of 'maximal margin hyperplanes.' In this context, we see that the constraint $\frac{1}{\|w\|_2} = C$ is fixing the *euclidean* distance between hyperplanes $w^T X + b = 1$ and $w^T X + b = -1$, i.e. fixing the 'margin', to be equal to C. Also, we see that the directional distance from X_i to the corresponding "separating" hyperplane is equal to $\frac{-y_i(w^T X_i + b) + 1}{\|w\|_2}$. The optimization problem above, therefore, fixes the margin between "separating" hyperplanes and minimizes the buffered probability of margin violations. If the classes are linearly separable with a margin of at least C, then the optimal objective is 0, meaning that "separating" hyperplanes are indeed separating. However, when classes are not separable with a margin of at least C, optimization problem (21) finds the number of worst classified objects such that the average of their directional distances to corresponding hyperplanes equals to 0. This number of worst classified objects is then minimized subject to the fixed margin size.

This interpretation, though, extends to any general norm within formulation (21). Using euclidean distance as our metric for measuring distances in $\mathbb{R}^n$, geometry tells us that the distance between hyperplanes is given by $\frac{2}{\|w\|_2}$ and that the directional distance from X_i to the corresponding "separating" hyperplane equals to $\frac{-y_i(w^T X_i + b) + 1}{\|w\|_2}$. But what if we were to use a different metric to measure distances within $\mathbb{R}^n$? For example, what if we were to say that the 'distance' between two points X_1, X_2 was $\|X_1 - X_2\|_1$ instead of $\|X_1 - X_2\|_2$? In this case, the distance between hyperplanes is given by $\frac{2}{\|w\|_\infty}$. This follows from the concept of the dual norm.

Denote by $\|\cdot\|^*$ the norm dual to the norm $\|\cdot\|$. In the general case, we know that if the 'distance' between two points X_1, X_2 is $\|X_1 - X_2\|^*$, then the distance between hyperplanes $w^T X + b = 1$ and $w^T X + b = -1$ is equal to $\frac{2}{\|w\|}$ and that the directional distance from X_i to the corresponding "separating" hyperplane equals to $\frac{-y_i(w^T X_i + b) + 1}{\|w\|}$. Thus, for the EC-SVM formulated with general norm $\|\cdot\|$, the constraint $\frac{1}{\|w\|} = C$ is fixing the 'margin' equal to C in $\mathbb{R}^n$ under the implied distance metric defined by the dual norm, $\|\cdot\|^*$.

Note that the problem above is equivalent to the EC-SVM, i.e. (19), having the same optimal objectives and (up to some scaling factor) equivalent optimal hyperplanes when

parameter C is the same for both problems. Therefore, the interpretation above is also valid for the EC-SVM. In particular, if a certain norm $\|\cdot\|$ is used in the optimization problem setting, then it is implied that distances between objects represent object similarities in a better fashion when measured according to the norm $\|\cdot\|^*$, dual to $\|\cdot\|$. The correspondence between C-SVM and EC-SVM, described further by Theorems 1 and 2, is not as direct as the correspondence between (21) and the EC-SVM. However, the presence of this correspondence should also imply that the use of $\|\cdot\|$ in optimization problem is closely related to the choice of norm $\|\cdot\|^*$ for the considered space.

Thus, as opposed to the C-SVM formulation, the EC-SVM formulation has a clear interpretation as hard margin separation problem for the case of non-separable classes.

6 Connecting the EC-SVM and C-SVM

In this section, we prove the equivalence of the EC-SVM and C-SVM when formulated with any general norm. Theorems 1 and 2 present the main result, providing a direct correspondence between parameter choices and optimal solutions. We emphasize the critical implication, which is that solving the C-SVM with any parameter $\hat{C} \geq 0$ is equivalent to solving the EC-SVM (i.e. minimizing bPOE) for some parameter $C \geq 0$.

Below be prove the equivalence of C-SVM and EC-SVM with two theorems. We leave the lengthy proofs to Appendix B. Theorem 1 begins with the assumption that one has solved the C-SVM and then provides the proper parameter value for which the EC-SVM will yield the same optimal hyperplane, up to a specific scaling factor, which we also provide. Theorem 2 is analogous to Theorem 1, but begins with the assumption that one has solved the EC-SVM. It should be noted that the theorems reference dual variables, which are discussed more explicitly (via KKT conditions) within the proofs of the theorems.

Theorem 1. *Assume that the data set is not linearly separable and suppose optimization problem (1) is formulated with any general norm $\|\cdot\|$ and some parameter $\hat{C} \geq 0$ and that it has optimal primal variables (w^*, b^*, ξ^*) and optimal dual variables (α^*, β^*). Then (19), formulated with corresponding norm and parameter $C = \frac{\hat{C}}{\sum_{i=1}^N \beta_i^*}$, will have optimal primal variables $(w = \mu w^*, b = \mu b^*, \xi = \mu \xi^*)$ and optimal dual variables $(\alpha = \alpha^*, \beta = \beta^*)$, where $\mu = \frac{\sum_{i=1}^N \beta_i^*}{\sum_{i=1}^N \beta_i^* - \hat{C}\|w^*\|} > 0$.*

Theorem 2. *Suppose optimization problem (19) is formulated with any general norm $\|\cdot\|$ and some parameter $C \geq 0$ and that it has optimal primal variables (w^*, b^*, ξ^*) and optimal dual variables (α^*, β^*). Then (1), formulated with corresponding norm and parameter $\hat{C} = C \sum_{i=1}^N \beta_i^*$, will have optimal primal variables $(w = \mu w^*, b = \mu b^*, \xi = \mu \xi^*)$ and optimal dual variables $(\alpha = \alpha^*, \beta = \beta^*)$, where $\mu = \frac{1}{1 + C\|w^*\|} > 0$.*

7 Presentation as Cohesive Structure

Here, we show exactly why it is appropriate to call formulation (19) the EC-SVM. Specifically, we show that the EC-SVM is equivalent to the Eν-SVM, producing the same set of optimal solutions. This follows directly from the fact that bPOE is the inverse of the superquantile. This fact helps solidify the idea that the EC-SVM is an **extension** of the C-SVM in the same way that the Eν-SVM is an extension of the ν-SVM. First, as was proved in Section 6, the optimal solution set produced by the C-SVM over $C \in [0, \infty)$ is contained in the optimal solution set produced by the EC-SVM over $C \in (-\infty, \infty)$. Second, the EC-SVM extends the allowable range of the C parameter to negative C-values. Thus, just as the Eν-SVM extends the parameter range and optimal solution set of the ν-SVM, the EC-SVM does the same for the C-SVM.

Proposition 3. *The EC-SVM and Eν-SVM, formulated with the same general norm, produce the same set of optimal hyperplanes.*

Proof. First, recall that the EC-SVM is equivalent to bPOE minimization. Second, recall that the Eν-SVM is equivalent to superquantile minimization. Using these two facts, the equivalence follows immediately from [4], which shows that (w, b) is a minimizer of bPOE at some threshold level if and only if (w, b) is a minimizer of the superquantile at some probability level. □

To help gather all of the results we have discussed, we can present them as a cohesive structure in the following table:

$\hat{C} \in [0, +\infty)$		$C \in (-\infty, +\infty)$	$=$	$-z$
$\uparrow$		$\uparrow$		$\uparrow$
C-SVM	$\subset$	EC-SVM	$\equiv$	$\min_{w,b}\ \bar{p}_z\left(-y(w^T X + b)\right)$
$\Updownarrow$		$\Updownarrow$		$\Updownarrow$
ν-SVM	$\subset$	Eν-SVM	$\equiv$	$\min_{w,b}\ \bar{q}_\alpha\left(-y(w^T X + b)\right)$
$\downarrow$		$\downarrow$		$\downarrow$
$\hat{\nu} \in (\nu_{\min}, \nu_{\max}]$		$\nu \in [0, 1]$	$=$	$1 - \alpha$

Key:

- $\Updownarrow$ Formulations generate the same set of optimal solutions.
- $\subset$ The right hand side formulation is an "extension" of the left hand formulation. (i.e. in the way that Eν-SVM is an extension of ν-SVM)
- $\equiv$ Formulations are objectively equivalent.
- $\uparrow$ or $\downarrow$ Arrow points to parameter values for the formulation.

8 Dual Formulations and Kernalization

Typically, the C-SVM is presented first in its primal form, then in its dual form, as the latter provides insights into selection of support vectors via dual variables, while additionally providing a quadratic optimization problem for solving the C-SVM (in the case of the L_2 regularization norm). This quadratic optimization problem also allows one to use the 'kernel trick,' utilizing the presence of dot products to introduce a non-linear mapping to a high dimensional feature space.

In this section, we present the dual formulations of the C-SVM, (1), and EC-SVM, (19), formulated with any general norm. We leave the derivation of these dual formulations to Appendix C. We use these formulations to enlighten our perspective in two ways. First, in conjunction with Theorems 1 and 2, we use the dual formulations to show that the optimal objective value of the C-SVM and EC-SVM coincide when yielding the same optimal hyperplane. This effectively yields the surprising result that the optimal value of the C-SVM objective function, divided by sample size, equals a probability level. Second, we use these dual formulations to present kernalization of the EC-SVM when formulated with the L_2 norm.

Assume the EC-SVM, (19), is formulated with some norm $\|\cdot\|$. Let $\|\cdot\|^*$ be the corresponding dual norm, then the dual formulation is as follows:

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^{N} \beta_i \\ s.t. \quad & \left\| \sum_{i=1}^{N} \beta_i y_i x_i \right\|^* \leq C \sum_{i=1}^{N} \beta_i, \\ & \sum_{i=1}^{N} \beta_i y_i = 0, \\ & 0 \leq \beta_i \leq 1, \quad \forall i \in \{1, \ldots, N\}. \end{aligned} \tag{22}$$

Assume the C-SVM, (1), is formulated with some norm $\|\cdot\|$. Let $\|\cdot\|^*$ be the corresponding dual norm, then the dual formulation is as follows:

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^{N} \beta_i \\ s.t. \quad & \left\| \sum_{i=1}^{N} \beta_i y_i x_i \right\|^* \leq C, \\ & \sum_{i=1}^{N} \beta_i y_i = 0, \\ & 0 \leq \beta_i \leq 1, \quad \forall i \in \{1, \ldots, N\}. \end{aligned} \tag{23}$$

Given the dual formulations, it follows immediately from Theorems 1 and 2 that the optimal dual objective solutions coincide when parameters are chosen so that the EC-SVM and C-SVM produce the same optimal hyperplane. Additionally, this result applies to the primal formulations via strong duality. Thus, since the EC-SVM objective, divided by sample size, equals a probability level (as seen in Property 6), we can conclude that the optimal objective value of the C-SVM primal formulation, divided by sample size, also equals a probability level.

8.1 Kernalization of EC-SVM

Here we briefly present the kernalization of the EC-SVM for the L_2 norm case, as it may be interesting to readers familiar with the C-SVM and its extensions to non-linear feature mappings. To effectively apply the kernel trick to the EC-SVM formulated with the L_2 norm [1], we need only to square the constraint

$$\left\| \sum_{i=1}^{N} \beta_i y_i x_i \right\|^* \leq C \sum_{i=1}^{N} \beta_i \ ,$$

forming the following quadratically constrained optimization problem (24), where $\phi(x_i)$ represents a non-linear kernel mapping of the i_{th} data vector.

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^{N} \beta_i \\ s.t. \quad & \sum_{i=1}^{N} \sum_{i=1}^{N} \beta_i \beta_j \left(y_i y_j \phi(x_i)^T \phi(x_j) - C^2 \right) \leq 0, \\ & \sum_{i=1}^{N} \beta_i y_i = 0, \\ & 0 \leq \beta_i \leq 1, \quad \forall i \in \{1, \ldots, N\}. \end{aligned} \tag{24}$$

9 Conclusion

In this paper we have introduced a new SVM formulation called the EC-SVM to help provide theoretical insights into the nature of the C-SVM, soft margin support vector classifier. Much like the Eν-SVM, this new formulation acts as an extension of the C-SVM. The main contribution of this paper, though, is not a new SVM formulation with computational or generalization benefits.

The main contribution of this paper is proof that soft margin support vector classification is equivalent to simple bPOE minimization. Additionally, we show that the C-SVM,

[1] Recall that the L_2 norm is self-dual.

EC-SVM, ν-SVM, and Eν-SVM fit nicely into the general framework of superquantile and bPOE minimization problems. This allows us to gain interesting and surprising insights, interpreting soft margin support vector optimization with newly developed statistical tools. For example, we were able to show that the C-parameter of the C-SVM has a statistical interpretation and that the optimal objective value, divided by sample size, equals a probability level.

Additionally, we were able to provide an interpretation of the EC-SVM as a hard-margin optimization problem, showing that the choice of regularization norm implies a metric used to define the margin. We were also able to provide dual formulations for both the C-SVM and EC-SVM formulated with any general regularization norm and provided a kernalized formulation for the EC-SVM.

In the broad scheme, we were able to show that the C-SVM formulation, derived traditionally from geometric intuition, can also be derived from purely statistical tools, with no geometric intuition involved. Specifically, we show that these statistical tools are superquantiles and the related bPOE.

Acknowledgment

Authors would like to thank Dr. Akiko Takeda, Dr. Anand Rangarajan, and Prof. R. Tyrrell Rockafellar, for the productive discussions and valuable comments.

This work was partially supported by the USA AFOSR grants: "Design and Redesign of Engineering Systems", FA9550-12-1-0427, and "New Developments in Uncertainty: Linking Risk Management, Reliability, Statistics and Stochastic Optimization", FA9550-11-1-0258. This material is based upon work supported by the Air Force Research Laboratory (AFRL) Mathematical Modeling and Optimization Institute.

A Ineffective variation of threshold levels

In application, it may be desirable that bPOE is minimized for different thresholds $z \in \mathbb{R}$ yielding a selection of optimal distributions $f(w_z^*, X)$, where $w_z^* = \arg\min \bar{p}_z(f(w, X))$ for some chosen value of threshold z. This way, one could do some type of model selection or analysis based upon the behavior of the optimal distribution over different thresholds. In doing so, one would expect to achieve different solutions for different threshold choices. As shown in the following propositions, the naive construction of formulation (12) combined with the positive homogeneity of $f(w, X)$ causes formulation (12) to achieve only two possible optimal solutions.

Proposition 4 and Corollary 1 show that for any threshold $z \leq 0$, formulation (12) becomes equivalent to

$$\min_{w \in \mathbb{R}^n} \quad \bar{p}_0(f(w, X)),$$

effectively yielding the solution for threshold $z = 0$. Proposition 5 shows that for any threshold $z > 0$, formulation (12) yields a trivial solution.

Proposition 4. *If $f(w, X)$ is PH w.r.t. w and minimizing bPOE at $z \leq 0$ yields*

$$1 - \alpha^* = \min_{w,\gamma<z} \quad \frac{E[f(w,X) - \gamma]^+}{z - \gamma},$$

with optimal solution vector (w^*, γ^*), then for any $a \geq 1$, $\bar{z} = az$ we have

$$1 - \alpha^* = \min_{w,\gamma<\bar{z}} \quad \frac{E[f(w,X) - \gamma]^+}{\bar{z} - \gamma},$$

with optimal solution vector $(aw^*, a\gamma^*)$.

Proof. Assume that for $z \leq 0$,

$$1 - \alpha^* = \min_{w\in\mathbb{R}^n} \quad \bar{p}_z(f(w,X)) = \min_{w\in\mathbb{R}^n,\gamma<z} \quad \frac{E[f(w,X) - \gamma]^+}{z - \gamma} = \frac{E[f(w^*,X) - \gamma^*]^+}{z - \gamma^*}.$$

This means that $\bar{p}_z(f(w^*, X)) \leq \bar{p}_z(f(w, X))$ for every $w \in \mathbb{R}^n$. Now, notice that for $\bar{z} = az$, where $a \geq 1$,

$$\begin{aligned}
\bar{p}_{\bar{z}}(f(aw^*, X)) &= \frac{E[f(aw^*,X) - a\gamma^*]^+}{\bar{z} - a\gamma^*} \\
&= \frac{a}{a}\frac{E[f(w^*,X) - \gamma^*]^+}{z - \gamma^*} \\
&= 1 - \alpha^* \\
&= \bar{p}_z(f(w^*, X)).
\end{aligned}$$

Since $\bar{p}_z(f(w, X))$ is a monotonically decreasing function w.r.t. z, we also know that $\bar{p}_z(f(w^*, X)) \leq \bar{p}_{\bar{z}}(f(w^*, X))$ for every $\bar{z} = az$, $a \geq 1$. Therefore, if $\min_{w\in\mathbb{R}^n} \bar{p}_z(f(w, X)) = 1 - \alpha^*$ at (w^*, γ^*), then for any $\bar{z} = az$, $a \geq 1$ we have that $\min_{w\in\mathbb{R}^n} \bar{p}_{\bar{z}}(f(w, X)) = 1 - \alpha^*$ at $(aw^*, a\gamma^*)$. □

Corollary 1. *Given Proposition 1, we can say that if $z \leq 0$, then*

$$\min_{w\in\mathbb{R}^n,\gamma<z} \quad \frac{E[f(w,X) - \gamma]^+}{z - \gamma} = \min_{w\in\mathbb{R}^n} \quad \bar{p}_0(f(w, X)).$$

Proof. Let (z_n) be a strictly decreasing sequence such that $z_0 < 0$ and $\lim_{n\to\infty} z_n = 0$. Proposition 4 implies that

$$\min_{w\in\mathbb{R}^n} \bar{p}_{z_0}(f(w, X)) = \min_{w\in\mathbb{R}^n} \bar{p}_{z_1}(f(w, X)) = ... = \min_{w\in\mathbb{R}^n} \bar{p}_0(f(w, X)).$$

Intuitively, this corollary simply follows from application of Proposition 4 to any $z < 0$ arbitrarily close to zero. □

Proposition 5. *If $z > 0$, then*

$$\min_{w\in\mathbb{R}^n,\gamma<z} \quad \frac{E[f(w,X)-\gamma]^+}{z-\gamma} = 0.$$

Proof. Objective is a non-negative function. If $z > 0$, then for $\gamma \in (0, z)$ the objective is 0, hence, optimal. □

B Proofs for Theorems 1 and 2

B.1 Theorem 1

Proof. To prove Theorem 1, we compare the KKT systems of (1) and (19) formulated with the same general norm, $\|\cdot\|$. We assume that the C-SVM with parameter $\hat{C} \geq 0$ yields optimal primal variables (w^*, b^*, ξ^*) and optimal dual variables (α^*, β^*). Thus, this is the same as assuming that $(w^*, b^*, \xi^*, \alpha^*, \beta^*)$ satisfies the following KKT system of (1):

$$\xi^* \geq 0, \tag{25a}$$
$$\beta^* \geq 0, \tag{25b}$$
$$\alpha^* \geq 0, \tag{25c}$$
$$-\alpha^* - \beta^* + 1 = 0, \tag{25d}$$
$$\xi_i^* \geq -y_i({w^*}^T x_i + b^*) + 1, \tag{25e}$$
$$0 \in -\hat{C}\partial\|w^*\| + \sum_{i=1}^N y_i\beta_i^* x_i, \tag{25f}$$
$$\sum_{i=1}^N -y_i\beta_i^* = 0, \tag{25g}$$
$$\alpha_i^*\xi_i^* = 0 = (1-\beta_i^*)\xi_i^*, \tag{25h}$$
$$\beta_i^*\left[-y_i({w^*}^T x_i + b^*) + 1 - \xi_i^*\right] = 0. \tag{25i}$$

Now, we show that with $\mu = \frac{\sum_{i=1}^N \beta_i^*}{\sum_{i=1}^N \beta_i^* - \hat{C}\|w^*\|}$, the variables $(\mu w^*, \mu b^*, \mu\xi^*, \alpha^*, \beta^*)$ satisfy the KKT system of (19). We then show that indeed $\mu > 0$ when the data set is not linearly

separable. The KKT system of (19) formulated with parameter $C \geq 0$ is as follows:

$$\xi \geq 0, \tag{26a}$$
$$\beta \geq 0, \tag{26b}$$
$$\alpha \geq 0, \tag{26c}$$
$$-\alpha - \beta + 1 = 0, \tag{26d}$$
$$\xi_i \geq -y_i(w^T x_i + b) + 1 + C\|w\|, \tag{26e}$$
$$0 \in -C\partial\|w\| \sum_{i=1}^{N} \beta_i + \sum_{i=1}^{N} y_i \beta_i x_i, \tag{26f}$$
$$\sum_{i=1}^{N} -y_i \beta_i = 0, \tag{26g}$$
$$\alpha_i \xi_i = 0 = (1 - \beta_i)\xi_i, \tag{26h}$$
$$\beta_i \left[-y_i(w^T x_i + b) + 1 + C\|w\| - \xi_i\right] = 0. \tag{26i}$$

When now show, one by one, that $\left(w = \mu w^*, b = \mu b^*, \xi = \mu \xi^*, \alpha = \alpha^*, \beta = \beta^*, C = \frac{\hat{C}}{\sum_{i=1}^{N} \beta_i^*}\right)$ satisfy all of these conditions and is thus a solution to the KKT system (26).

1. $\xi \geq 0$: True, since $\xi = \mu\xi^*$, $\xi^* \geq 0$, $\mu \geq 0$.
2. $\beta \geq 0$: True, since $\beta = \beta^* \geq 0$.
3. $\alpha \geq 0$: True, since $\alpha = \alpha^*$.
4. $-\alpha - \beta + 1 = 0$: True, since $-\alpha - \beta + 1 = -\alpha^* - \beta^* + 1 = 0$.
5. $\xi_i \geq -y_i(w^T x_i + b) + 1 + C\|w\| \longrightarrow \mu\xi_i^* \geq -y_i(w^{*T} x_i + b^*)\mu + 1 + \mu C\|w^*\|$
   $\longrightarrow \xi^* \geq -y_i(w^{*T} x_i + b^*) + \left(\frac{1}{\mu} + C\|w^*\|\right) = -y_i(w^{*T} x_i + b^*) + \left(1 - \frac{\hat{C}}{\sum_{i=1}^{N} \beta^*}\|w^*\| + \frac{\hat{C}}{\sum_{i=1}^{N} \beta^*}\|w^*\|\right) =$
   $-y_i(w^{*T} x_i + b^*) + 1$: True, following from (25e).
6. $0 \in -C\partial\|w\| \sum_{i=1}^{N} \beta_i + \sum_{i=1}^{N} y_i \beta_i x_i$: True, following from $C\partial\|w\| \sum_{i=1}^{N} \beta_i = \hat{C}\partial\|w^*\|$,
   $\sum_{i=1}^{N} y_i \beta_i x_i = \sum_{i=1}^{N} y_i \beta_i^* x_i$, and (25f).
7. $\sum_{i=1}^{N} -y_i \beta_i = 0$: True, since $\beta = \beta^*$ and (25g).
8. $\alpha_i \xi_i = 0$: True, since $\alpha = \alpha^*$, $\xi = \mu\xi^*$, $\mu > 0$, and (25h).
9. $\beta_i \left[-y_i(w^T x_i + b) + 1 + C\|w\| - \xi_i\right] = 0 \iff \beta_i \xi_i = -\beta_i y_i(w^T x_i + b) + \beta_i + \beta_i C\|w\|$.
   Notice then that (26h) $\implies \beta_i \xi_i = \xi_i$. This gives us $\xi_i = \mu\xi^* = -\mu\beta_i^* y_i(w^T x_i + b^*) +$

$\beta_i^* + \beta_i^* C\|w^*\| = \beta_i^* + \beta_i^* \frac{\hat{C}\mu}{\sum_{i=1}^N \beta_i^*}\|w^*\| + \mu\left[\xi_i^* - \beta_i^*\right] \iff \beta_i^*\left[1 + \frac{\hat{C}\mu}{\sum_{i=1}^N \beta_i^*}\|w^*\| - \mu\right] = 0$
$\iff 1 + \frac{\hat{C}\|w^*\|}{\sum_{i=1}^N \beta_i^* - \hat{C}\|w^*\|} - \frac{\sum_{i=1}^N \beta_i^*}{\sum_{i=1}^N \beta_i^* - \hat{C}\|w^*\|} = 0$. Clearly, the last equality is true.

Now we show that for non-linearly separable data sets, $\mu > 0$. We use the KKT system (25) to form the dual via the Lagrangian. We then show that $\sum_{i=1}^N \beta_i^* > \hat{C}\|w^*\|$, which proves that $\mu > 0$.

We first have that the Lagrangian of (1) is the following:

$$L(w, b, \xi, \alpha, \beta) = \hat{C}\|w\| + \sum_{i=1}^N \xi_i + \sum_{i=1}^N \beta_i\left[-y_i(w^T x_i + b) + 1 - \xi_i\right] - \sum_{i=1}^N \alpha_i \xi_i$$

Using (25d,g), we can then simplify, also maximizing w.r.t. the dual variables and minimizing w.r.t. the primal variables, to form the following dual. Here, the constraints are implied by (25d,g).

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^N \beta_i + \left(\inf_{w,b} \hat{C}\|w\| + \sum_{i=1}^N \beta_i\left[-y_i(w^T x_i + b)\right]\right) \\ s.t. \quad & 0 \le \beta \le 1. \\ & \sum_{i=1}^N y_i \beta_i = 0. \end{aligned} \tag{27}$$

Notice that since $w = 0$ is a feasible solution, we know that

$$\inf_{w,b} \quad \hat{C}\|w\| + \sum_{i=1}^N \beta_i\left[-y_i(w^T x_i + b)\right] \le 0.$$

Finally, noting that with the optimal variables assumed to be $(w^*, b^*, \xi^*, \alpha^*, \beta^*)$, we see that

$$\sum_{i=1}^N \beta_i^* + \left(\inf_{w,b} \quad \hat{C}\|w\| + \sum_{i=1}^N \beta_i^*\left[-y_i(w^T x_i + b)\right]\right) = \hat{C}\|w^*\| + \sum_{i=1}^N \xi^*.$$

But since the data set is not linearly separable $\sum_{i=1}^N \xi^* > 0$. This implies that $\sum_{i=1}^N \beta_i^* > \hat{C}\|w^*\|$, which shows that $\mu = \frac{\sum_{i=1}^N \beta_i^*}{\sum_{i=1}^N \beta_i^* - \hat{C}\|w^*\|} > 0$. □

B.2 Theorem 2

Proof. To prove Theorem 2, we compare the KKT systems of (1) and (19) formulated with the same general norm, $\|\cdot\|$. We assume that the EC-SVM with parameter $C \ge 0$ yields

optimal primal variables (w^*, b^*, ξ^*) and optimal dual variables (α^*, β^*). Thus, this is the same as assuming that $(w^*, b^*, \xi^*, \alpha^*, \beta^*)$ satisfies the following KKT system of (19):

$$\xi^* \geq 0, \tag{28a}$$
$$\beta^* \geq 0, \tag{28b}$$
$$\alpha^* \geq 0, \tag{28c}$$
$$-\alpha^* - \beta^* + 1 = 0, \tag{28d}$$
$$\xi_i^* \geq -y_i(w^{*T}x_i + b^*) + 1 + C\|w^*\|, \tag{28e}$$
$$0 \in -C\partial\|w^*\| \sum_{i=1}^{N} \beta_i^* + \sum_{i=1}^{N} y_i\beta_i^* x_i, \tag{28f}$$
$$\sum_{i=1}^{N} -y_i\beta_i^* = 0, \tag{28g}$$
$$\alpha_i^*\xi_i^* = 0 = (1-\beta_i^*)\xi_i^*, \tag{28h}$$
$$\beta_i^* \left[-y_i(w^{*T}x_i + b^*) + 1 + C\|w^*\| - \xi_i^*\right] = 0. \tag{28i}$$

Now, we show that with $\mu = \frac{1}{1+C\|w^*\|} > 0$, the variables $(\mu w^*, \mu b^*, \xi^*, \alpha^*, \beta^*)$ satisfy the KKT system of (1). We then show that indeed $\mu > 0$ when the data set is not linearly separable. The KKT system of (1) formulated with parameter $\hat{C} \geq 0$ is as follows:

$$\xi \geq 0, \tag{29a}$$
$$\beta \geq 0, \tag{29b}$$
$$\alpha \geq 0, \tag{29c}$$
$$-\alpha - \beta + 1 = 0, \tag{29d}$$
$$\xi_i \geq -y_i(w^T x_i + b) + 1, \tag{29e}$$
$$0 \in -\hat{C}\partial\|w\| + \sum_{i=1}^{N} y_i\beta_i x_i, \tag{29f}$$
$$\sum_{i=1}^{N} -y_i\beta_i = 0, \tag{29g}$$
$$\alpha_i\xi_i = 0 = (1-\beta_i)\xi_i, \tag{29h}$$
$$\beta_i \left[-y_i(w^T x_i + b) + 1 - \xi_i\right] = 0. \tag{29i}$$

When now show, one by one, that $(w = \mu w^*, b = \mu b^*, \xi = \mu\xi^*, \alpha = \alpha^*, \beta = \beta^*, \hat{C} = C\sum_{i=1}^{N}\beta_i)$ satisfy all of these conditions and is thus a solution to the KKT system (29):

1. $\xi \geq 0$: True, since $\xi = \mu\xi^*$, $\xi^* \geq 0$, $\mu \geq 0$.

2. $\beta \geq 0$: True, since $\beta = \beta^* \geq 0$.

3. $\alpha \geq 0$: True, since $\alpha = \alpha^*$.

4. $-\alpha - \beta + 1 = 0$: True, since $-\alpha - \beta + 1 = -\alpha^* - \beta^* + 1 = 0$.

5. $\xi_i \geq -y_i(w^T x_i + b) + 1 \longrightarrow \mu\xi_i^* \geq -y_i(w^{*T} x_i + b^*)\mu + 1 \iff \xi^* \geq -y_i(w^{*T} x_i + b^*) + (1 + C\|w^*\|)$: True, following from (28e).

6. $0 \in -\hat{C}\partial\|w^*\| + \sum_{i=1}^N y_i\beta_i x_i$: True, following from $\hat{C}\partial\|w^*\| = C\partial\|w\| \sum_{i=1}^N \beta_i$, $\sum_{i=1}^N y_i\beta_i x_i = \sum_{i=1}^N y_i\beta_i^* x_i$, and (28f).

7. $\sum_{i=1}^N -y_i\beta_i = 0$: True, since $\beta = \beta^*$ and (28g).

8. $\alpha_i\xi_i = 0$: True, since $\alpha = \alpha^*$, $\xi = \mu\xi^*$, $\mu > 0$, and (28h).

9. $\xi\mu^* = \xi = \beta_i + \mu\left[-\beta_i y_i(w^{*T} x_i + b^*)\right] = \beta_i + \mu\left[\xi^* - \beta - \beta C\|w^*\|\right] = \mu\xi^* + \beta\left(1 - \frac{1}{1+C\|w^*\|} - \frac{C\|w^*\|}{1+C\|w^*\|}\right) = \mu\xi^*$: Clearly, the last equality is true.

Now we point out that $\mu > 0$. This follows immediately from the assumption that $C \geq 0$ and the fact that $\|w^*\| \geq 0$. □

C Derivation of EC-SVM Dual Formulation

The dual of the EC-SVM is formulated as follows, via the Lagrangian:

$$
\begin{aligned}
&\max_{\substack{\alpha\geq 0\\ \beta\geq 0}} \quad \min_{w,b} \quad \sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\beta_i\left[-y_i(w^Tx_i+b)+1-\xi_i+C\|w\|\right] - \sum_{i=1}^{N}\alpha_i\xi_i \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i=0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{w} \quad C\|w\|\sum_{i=1}^{N}\beta_i + \sum_{i=1}^{N}-\beta_i y_i w^Tx_i\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i=0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{w\neq 0,a\geq 0} \quad aC\sum_{i=1}^{N}\beta_i + \frac{a\sum_{i=1}^{N}-\beta_i y_i w^Tx_i}{\|w\|}\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i=0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{a\geq 0} \quad aC\sum_{i=1}^{N}\beta_i + a\min_{w\neq 0}\frac{\sum_{i=1}^{N}-\beta_i y_i w^Tx_i}{\|w\|}\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i=0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{a\geq 0} aC\sum_{i=1}^{N}\beta_i - a\min_{w\neq 0} \quad \frac{\sum_{i=1}^{N}\beta_i y_i w^Tx_i}{\|-w\|}\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i=0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{a\geq 0} \quad aC\sum_{i=1}^{N}\beta_i - a\left\|\sum_{i=1}^{N}-\beta_i y_i x_i\right\|^*\right] \\
&= \max_{\beta} \quad \sum_{i=1}^{N}\beta_i \\
&\qquad s.t. \quad \left\|\sum_{i=1}^{N}\beta_i y_i x_i\right\|^* \leq C\sum_{i=1}^{N}\beta_i, \\
&\qquad\qquad \sum_{i=1}^{N}\beta_i y_i = 0, \\
&\qquad\qquad 0\leq\beta_i\leq 1, \quad \forall i\in\{1,\ldots,N\}.
\end{aligned}
$$

D Derivation of C-SVM Dual Formulation

The dual of the C-SVM is formulated as follows, via the Lagrangian:

$$
\begin{aligned}
&\max_{\substack{\alpha\geq 0\\ \beta\geq 0}} \quad \min_{w,b} \quad \sum_{i=1}^{N}\xi_i + C\|w\| + \sum_{i=1}^{N}\beta_i\left[-y_i(w^T x_i + b) + 1 - \xi_i\right] - \sum_{i=1}^{N}\alpha_i\xi_i \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i = 0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{w} \quad C\|w\| + \sum_{i=1}^{N} -\beta_i y_i w^T x_i\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i = 0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{w\neq 0, a\geq 0} \quad aC + \frac{a\sum_{i=1}^{N} -\beta_i y_i w^T x_i}{\|w\|}\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i = 0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{a\geq 0} \quad aC + a\min_{w\neq 0}\frac{\sum_{i=1}^{N} -\beta_i y_i w^T x_i}{\|w\|}\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i = 0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{a\geq 0} \quad aC - a\min_{w\neq 0}\frac{\sum_{i=1}^{N} \beta_i y_i w^T x_i}{\|-w\|}\right] \\
&= \max_{\substack{0\leq\beta\leq 1\\ \sum_{i=1}^{N}\beta_i y_i = 0}} \quad \sum_{i=1}^{N}\beta_i + \left[\min_{a\geq 0} \quad aC - a\left\|\sum_{i=1}^{N} -\beta_i y_i x_i\right\|^*\right] \\
&= \max_{\beta} \quad \sum_{i=1}^{N}\beta_i \\
&\qquad s.t. \quad \left\|\sum_{i=1}^{N}\beta_i y_i x_i\right\|^* \leq C, \\
&\qquad\qquad \sum_{i=1}^{N}\beta_i y_i = 0, \\
&\qquad\qquad 0\leq\beta_i\leq 1, \quad \forall i\in\{1,\ldots,N\}.
\end{aligned}
$$

References

[1] Barbero, A., Takeda, A., Lopez, J. Geometric Intuition and Algorithms for Ev-SVM Journal of Machine Learning Research 16 (2015) 323-369

[2] Chang, C. C., and Lin, C. J. (2001) Training v-Support Vector Classifiers: Theory and Algorithms Neural computation, 13(9), 2119-2147

[3] Davis, J.R., Uryasev S. (2014). Analysis of Hurricane Damage using Buffered Probability of Exceedance Research Report 2014-4, ISE Dept., University of Florida.

[4] Mafusalov, A. and S. Uryasev. (2014) Buffered Probability of Exceedance: Mathematical Properties and Optimization Algorithms Research Report 2014-1, ISE Dept., University of Florida, October 2014

[5] Norton M., Mafusalov, A, Uryasev S. (2015) Cardinality of Upper Average and Application to Network Optimization Research Report 2015-1, ISE Dept., University of Florida

[6] Norton M., Uryasev S. (2014) Maximization of AUC and Buffered AUC in Classification Research Report 2014-2, ISE Dept., University of Florida, October 2014

[7] Prez-Cruz, F., Weston, J., Herrmann, D. J. L., and Scholkopf, B. (2003) Extension Of The ν-SVM Range For Classification NATO SCIENCE SERIES SUB SERIES III COMPUTER AND SYSTEMS SCIENCES, 190, 179-196.

[8] Rockafellar, R. T., and Uryasev, S. (2002) Conditional value-at-risk for general loss distributions Journal of Banking and Finance (2002), 1443-1471.

[9] Rockafellar, R.T. (2009). Safeguarding Strategies in Risky Optimization. Presentation at the International Workshop on Engineering Risk Control and Optimization, Gainesville, FL, February, 2009.

[10] Rockafellar R.T., Royset J.O. (2010). On Buffered Failure Probability in Design and Optimization of Structures. Reliability Engineering & System Safety, Vol. 95, 499-510.

[11] Schlkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000) New Support Vector Algorithms Neural computation, 12(5), 1207-1245

[12] Takeda, A., and Sugiyama, M. (2008) v-Support Vector Machine as Conditional Value-At-Risk Minimization In Proceedings of the 25th international conference on Machine learning (pp. 1056-1063). ACM.

[13] Uryasev, S. (2014) Buffered Probability of Exceedance and Buffered Service Level: Definitions and Properties. Research Report 2014-3, ISE Dept., University of Florida