Journal of Machine Learning Research 18 (2017) 1-43 Submitted 11/15; Revised 6/17; Published 7/17

Soft Margin Support Vector Classification as Buffered Probability Minimization

Matthew Norton MDNORTO@GMAIL.COM
Alexander Mafusalov SASHA.MAFUSALOV@GMAIL.COM
Stan Uryasev URYASEV@UFL.EDU
Risk Management and Financial Engineering Lab
Department of Industrial and Systems Engineering
University of Florida
Gainesville, FL 32611, USA.

Editor: Corinna Cortes

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), a recently introduced characterization of uncertainty. To show this, we introduce a new SVM formulation, called the EC-SVM, which is derived from a simple bPOE minimization problem that is easy to interpret with a meaningful free parameter, optimal objective value, and probabilistic derivation. Over the range of its free parameter, the EC-SVM has both a convex and non-convex case which we connect to existing SVM formulations. We first show that the C-SVM, formulated with any regularization norm, is equivalent to the convex EC-SVM. Similarly, we show that the Eν-SVM is equivalent to the EC-SVM over its entire parameter range, which includes both the convex and non-convex case. 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. We also show that the EC-SVM can easily be cast as a robust optimization problem, where bPOE is minimized with data lying in a fixed uncertainty set. This reformulation allows us to clearly differentiate between the convex and non-convex case, with convexity associated with pessimistic views of uncertainty and non-convexity associated with optimistic views of uncertainty. Finally, we address some practical considerations. First, we show that these new insights can assist in making parameter selection more efficient. Second, we discuss optimization approaches for solving the EC-SVM. Third, we address the issue of generalization, providing generalization bounds for both bPOE and misclassification rate.

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

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 Takeda and Sugiyama (2008), 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 Rockafellar and Uryasev (2002) as means for dealing with optimization of quantiles. Utilizing the popular calculation formula for superquantiles, Takeda and Sugiyama (2008) 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 Rockafellar (2009) and further studied in Rockafellar and Royset (2010). This generalization, recently studied by Mafusalov and Uryasev (2015); Norton and Uryasev (2014); Norton et al. (2015); Davis and Uryasev (2014); Uryasev (2014), 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 related 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 and Sugiyama (2008). 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 a novel interpretation of the EC-SVM as a robust optimization (RO) problem, where bPOE is minimized with data points lying in a fixed uncertainty set. With the EC-SVM having both a convex and non-convex case, depending on the value of the free parameter, we find that the RO representation reveals a unique interpretation to help distinguish these cases. Specifically, the RO representation reveals that the convex case corresponds to bPOE minimization with worst-case *pessimistic* data uncertainty, while the non-convex case corresponds to bPOE minimization with best-case *optimistic* data uncertainty. The RO representation also provides other insights. For the convex case, the RO representation is similar to existing equivalences for the C-SVM presented in Xu et al. (2009), but with a much simpler proof, definition of uncertainty set, and correspondence between equivalent parameter values. For the non-convex case, this new representation suggests an efficient alternating minimization method for finding a local minimum. Furthermore, the RO representation reveals that this seemingly heuristic optimization method is related to the DC Algorithm, a popular algorithm for finding the local minimum of DC (difference of convex functions) optimization problems.

Finally, we consider some practical implications of these new theoretical insights. We show that the new interpretation of the C tradeoff parameter suggests a range that should be used to select the C parameter. This helps to improve grid selection for cross validation so that one can partially avoid solving the EC-SVM for values of C that yield trivial or redundant optimal solutions. Furthermore, using the RO representation of the EC-SVM, we show that there may be situations where prior knowledge about data uncertainty suggests, at-best, a fixed C or, at-worst, much tighter bounds for the range of interesting C values for cross-validation.

We also address the practical issue of generalization. In classification, generalization bounds are typically provided for misclassification rate because classification algorithms are viewed as attempts to find classifiers which minimize this performance metric. In this regard, we utilize results from Takeda and Sugiyama (2008); Schölkopf et al. (2000) to show that generalization bounds for misclassification rate of EC-SVM classifiers can be posed in terms of empirical estimates of bPOE, the quantile, and the superquantile. However, with the new insight that the EC-SVM and C-SVM directly minimize bPOE, we also provide generalization bounds for this metric. We apply the stability arguments of Bousquet and Elisseeff (2002) to provide tight generalization bounds on the true bPOE of EC-SVM and C-SVM classifiers and show that bPOE threshold plays an important role in these bounds.

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 and Sugiyama (2008), 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 of 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.

Section 9 shows that the EC-SVM has a novel interpretation as a RO problem, with a distinction between the convex and non-convex cases. Section 10.1 discusses the issue of parameter selection. Section 10.2 discusses optimization of the EC-SVM. Section 10.3 provides generalization bounds for misclassification rate and bPOE.

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 from Cortes and Vapnik (1995), 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.[1] 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 by Schölkopf et al. (2000) 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 Chang and Lin (2001). 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.

[1] Often, the squared norm is utilized. We use non-squared for ease of presentation. Many of the ideas in this paper are also valid for the squared case, however, the proofs are more involved and the correspondence between formulations is less straight-forward.

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 Schölkopf et al. (2000) 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:*

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

- *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:*

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

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 Chang and Lin (2001), it was shown that the ν-parameter has a limited range. Specifically, a variant of Property 2 is proved in Chang and Lin (2001).

Property 2

- *There exists a minimum and maximum value such that $\nu \in (\nu_{\min}, \nu_{\max}]$ 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, Pérez-Cruz et al. (2003) 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 Pérez-Cruz et al. (2003); Chang and Lin (2001) 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 Rockafellar and Uryasev (2002). We then review the results of Takeda and Sugiyama (2008), 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 (2002) 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 Takeda and Sugiyama (2008), the meaning of the ν-parameter was solidified by showing that the Eν-SVM, (3), is equivalent to superquantile minimization. Specifically, they proved a variant of Property 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:*

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

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

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, $p_z(Z) = P(Z > z)$, 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 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. Mafusalov and Uryasev (2015) mainly work with so called Lower bPOE while Norton and Uryasev (2014) work with so called Upper bPOE. These definitions do not differ dramatically and a discussion of these differences is 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,*

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

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, Norton and Uryasev (2014) provide 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,*

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

It is also important to note that formula (9) has the following property, Property 4, proved in Mafusalov and Uryasev (2015) and Norton and Uryasev (2014). This property will become important in later sections when we begin to interpret the EC-SVM. Note that for Property 4, we must distinguish between the *lower* quantile $q_\alpha(Z)$ and *upper* quantile $q_\alpha^+(Z) = \inf\{z | P(Z \leq z) > \alpha\}$. However, the difference in these quantities will likely be small or zero. For continuously distributed Z, we will have $q_\alpha(Z) = q_\alpha^+(Z)$ or if Z is discretely distributed with N equally probably events, and N is reasonably large, the difference $q_\alpha^+(Z) - q_\alpha(Z)$ will likely be small, possibly zero.

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:*

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

where $[q_{\alpha^*}(Z), q_{\alpha^*}^+(Z)]$ is the entire set of minimizers.

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

Norton and Uryasev (2014) considered the following optimization setup to demonstrate the ease with which bPOE can be minimized directly. Assume we have a real valued positive homogenous (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 $z \in \mathbb{R}$. We would like to solve the 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 Norton and Uryasev (2014), though, limit consideration to only a threshold of $z = 0$, in which case formulation (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

$$\begin{array}{cc} \min\limits_{w\in\mathbb{R}^n,\gamma<z} & \dfrac{E[f(w,X)-\gamma]^+}{z-\gamma} \\ s.t. & \|w\|=1 \end{array} \quad \equiv \quad \min_{w\in\mathbb{R}^n} \bar{p}_z\left(\frac{f(w,X)}{\|w\|}\right). \tag{14}$$

Furthermore, the following Proposition 2 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 with respect to 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

$$\begin{array}{cc} \min\limits_{w\in\mathbb{R}^n,\gamma<z} & E\left[f\left(\dfrac{w}{z-\gamma},X\right)-\left(\dfrac{\gamma}{z-\gamma}\right)\right]^+ \\ s.t. & \|w\|=1. \end{array} \tag{16}$$

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

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

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

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

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 2 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^*:

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

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:

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

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 Occurrence 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 description, 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.*

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

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

5.3.2 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 (for the full derivation, see Appendix E):

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

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, (19), having the same optimal objective and (up to some scaling factor) equivalent optimal hyperplane 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 dual norm. 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 Mafusalov and Uryasev (2015), 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:

$$\begin{array}{ccccc}
\hat{C} \in [0, +\infty) & & C \in (-\infty, +\infty) & = & -z \\
\uparrow & & \uparrow & & \uparrow \\
\hline
\text{C-SVM} & \subset & \text{EC-SVM} & \equiv & \min\limits_{w,b} \; \bar{p}_z\left(-y(w^T X + b)\right) \\
\Updownarrow & & \Updownarrow & & \Updownarrow \\
\nu\text{-SVM} & \subset & \text{E}\nu\text{-SVM} & \equiv & \min\limits_{w,b} \; \bar{q}_\alpha\left(-y(w^T X + b)\right) \\
\hline
\downarrow & & \downarrow & & \downarrow \\
\hat{\nu} \in (\nu_{\min}, \nu_{\max}] & & \nu \in [0, 1] & = & 1 - \alpha
\end{array}$$

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 $\|\cdot\|$ and $C \geq 0$. Let $\|\cdot\|^*$ be the corresponding dual norm, then the dual formulation is as follows:

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

Assume the C-SVM, (1), is formulated with some norm $\|\cdot\|$ and $C \geq 0$. Let $\|\cdot\|^*$ be the corresponding dual norm, then the dual formulation is as follows:

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

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 various methods for utilizing the kernel trick with the EC-SVM. First, the most obvious way is for the L_2 norm case and the dual EC-SVM formulation. To effectively apply the kernel trick[2], we need only to square the constraint in the dual formulation

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

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

forming the following convex quadratically constrained quadratic program (QCQP) (24), where $\phi(X_i)$ represents a non-linear kernel mapping of the i_{th} data vector. Note that this can be efficiently solved with most convex optimization software.

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

The application of the kernel trick, though, is not limited to the dual formulation. Mimicking the work of Chapelle (2007), we can apply the kernel in the primal formulated with the L_2 norm by using the representers theorem of Schölkopf et al. (2001). Since the representers theorem applies to the C-SVM, it indeed can be applied to the equivalent convex EC-SVM.[3] Thus, we know that in the high dimensional feature space we can represent the optimal solution of the convex EC-SVM as $w = \sum_i \theta_i y_i \phi(X_i)$. Therefore, using this representation, we can write the primal EC-SVM with $C \geq 0$ and the L_2 norm as the following unconstrained non-smooth convex program, where $K_{ij} = \phi(X_i)^T \phi(X_j)$.

$$\min_{\theta,b} \sum_{i=1}^{N} [-y_i(\sum_j \theta_j y_j K_{ij} + b) + C \left(\sum_i \sum_j \theta_i \theta_j y_i y_j K_{ij} \right)^{\frac{1}{2}} + 1]^+ \ . \tag{25}$$

Furthermore, this can be reformulated as the following convex QCQP.

$$\begin{aligned} \min_{\theta,b,\xi,t} \quad & \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(\sum_j \theta_j y_j K_{ij} + b) + Ct + 1, \quad \forall i \in \{1,\dots,N\}. \\ & \xi \geq 0 \\ & t^2 \geq \sum_i \sum_j \theta_i \theta_j y_i y_j K_{ij}. \end{aligned} \tag{26}$$

So far, we have discussed the case of the L_2 norm and $C \geq 0$. Here, for the interest of the reader, we briefly mention how to apply a standard technique which allows approximate kernelization of the EC-SVM with any norm and $C \in \mathbb{R}$. Here, we explicitly map the data vectors into the high dimensional feature space by approximating the kernel map. Let $K \in \mathbb{R}^{N \times N}$ be the positive semi-definite (PSD) matrix such that $K_{ij} = \phi(X_i)^T \phi(X_j)$ and, with abuse of notation, let $X = [X_1, ..., X_N]$ be the matrix of data vectors. Let

[3] In other words, because the EC-SVM and C-SVM produce the same set of optimal solutions for $C \geq 0$, we can say that their optimal solutions can be represented in the same way.

$V = [V_1, ..., V_N]$ and $\Lambda = diag\{\lambda_1, ..., \lambda_n\}$ be the matrix of eigenvectors and associated eigenvalues of K with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. With K being symmetric PSD, we then have the eigendecomposition,

$$K = \phi(X)^T \phi(X) = V \Lambda V^T = (\Lambda^{\frac{1}{2}} V^T)^T (\Lambda^{\frac{1}{2}} V^T) \ .$$

Thus, we can approximate $\phi(X)$ by taking the top d eigenvectors and eigenvalues to get

$$\hat{X} = diag\{\lambda_1, ..., \lambda_d\}^{\frac{1}{2}} [V_1, ..., V_d]^T \approx \Lambda^{\frac{1}{2}} V^T = \phi(X) \ .$$

Then, we can simply use $\hat{X}$ as our data in the linear EC-SVM, (19), with any norm or parameter value and solve using the standard methods that will be discussed in Section 10.2.

9. Interpretation as Robust Optimization, Pessimistic vs. Optimistic Uncertainty

In this section, we present an equivalent representation of the convex and non-convex case of the EC-SVM as robust optimization (RO) problems. This equivalence is enlightening for both the convex and non-convex case. In general, we find that the RO representation reveals a unique interpretation, showing that the convex case corresponds to bPOE minimization with worst-case *pessimistic* data uncertainty, while the non-convex case corresponds to bPOE minimization with best-case *optimistic* data uncertainty. For the convex case, the RO representation is similar to existing equivalences for the C-SVM in Xu et al. (2009), but with a much simpler proof, definition of uncertainty set, and correspondence between equivalent parameter values. For the non-convex case, this new representation reveals a connection with the Total Support Vector Classifier (T-SVC) proposed in Bi and Zhang (2005) and suggests an efficient heuristic for the optimization. Furthermore, the RO representation reveals that this seemingly heuristic optimization method, which was shown to work well in the T-SVC case, is related to DCA, a popular algorithm for locally minimizing DC (difference of convex) functions.

9.1 Convex Case and Pessimism

For the convex case of the EC-SVM, (19), when we have $C \geq 0$, it is simple to show that it is equivalent to (27) and (28). Notice that we are minimizing bPOE, as we did before, but with threshold equal to zero and data subject to *pessimistic* disturbances δ_i. We call this *pessimism*, because we are maximizing w.r.t. the disturbance and looking at the worst case. This is the traditional viewpoint of robust optimization, leading to a special case of the Wald's minimax model.

$$\begin{aligned} \min_{w,b} \quad \max_{\delta_i} \quad & \bar{p}_0\left(-y(w^T(X+\delta)+b)\right) \\ s.t. \quad & \|\delta_i\|^* \leq C \ , i = 1, ..., N \end{aligned} \tag{27}$$

$$\begin{aligned} \min_{w,b} \quad \max_{\delta_i} \quad & \sum_i [-y_i(w^T(X_i+\delta_i)+b)+1]^+ \\ s.t. \quad & \|\delta_i\|^* \leq C, \ i = 1, ..., N \end{aligned} \tag{28}$$

To see the equivalence between (19) and (28), one only needs to observe that for any $w \in \mathbb{R}^n$, if $C \geq 0$, we have that

$$C\|w\| = C \max_{\|\delta\|^* \leq 1} w^T\delta = \max_{\|\delta\|^* \leq C} w^T\delta = \max_{\|\delta\|^* \leq C} -yw^T\delta \ .$$

Furthermore, to see the equivalence between (28) and (27), a simple application of the bPOE formula (9) and multiplication by N transforms (27) into (28).

This provides us with three things. First, we see that the bPOE threshold can be interpreted as controlling the size of the uncertainty set. Following from the transformation of (19) to (28) via the simple algebra, we see that this correspondence is exact, meaning that a threshold of C in the EC-SVM implies an uncertainty set with "size" C in (28) (i.e. optimal solutions are the same when parameters are the same). Second, we see that regularization with norm $\|\cdot\|$ implies uncertainty w.r.t. the corresponding dual norm $\|\cdot\|^*$. Third, we see that the convex case can be interpreted as pessimistic, or risk averse, following the traditional RO framework.

9.2 Nonconvex Case and Optimism

For the non-convex case, we find a similar robust reformulation, but instead of pessimistic disturbances we have *optimistic* disturbances. Specifically, for the EC-SVM (19), when we have $C < 0$, it is simple to show that it is equivalent to (29) and (30). Notice again that we are minimizing bPOE, as we did before, but with threshold equal to zero and data subject to *optimistic* disturbances δ_i. We call this *optimistic*, because we are minimizing w.r.t. the disturbance and looking at the best case.

$$\begin{aligned} \min_{w,b,\delta_i} \quad & \bar{p}_0\left(-y(w^T(X+\delta)+b)\right) \\ s.t. \quad & \|\delta_i\|^* \leq -C \ , i = 1, ..., N \end{aligned} \tag{29}$$

$$\begin{aligned} \min_{w,b,\delta_i} \quad & \sum_i [-y_i(w^T(X_i+\delta_i)+b)+1]^+ \\ s.t. \quad & \|\delta_i\|^* \leq -C \ , i = 1, ..., N \end{aligned} \tag{30}$$

To see the equivalence between (19) and (30), one only needs to observe that for any $w \in \mathbb{R}^n$, if $C < 0$, we have that

$$\begin{aligned} C\|w\| &= C \max_{\|\delta\|^* \leq 1} w^T\delta \\ &= -\max_{\|\delta\|^* \leq 1} -Cw^T\delta \\ &= -\max_{\|\delta\|^* \leq -C} w^T\delta \\ &= \min_{\|\delta\|^* \leq -C} w^T(-\delta) \\ &= \min_{\|\delta\|^* \leq -C} w^T\delta \quad = \min_{\|\delta\|^* \leq -C} -yw^T\delta \ . \end{aligned}$$

Furthermore, to see the equivalence between (30) and (29), a simple application of the bPOE formula (9) and multiplication by N transforms (29) into (30).

This formulation is enlightening for two reasons. First, as already mentioned, we see that the non-convex EC-SVM can be viewed from the RO lens as minimizing bPOE with an optimistic view of the data set. The optimistic perspective makes sense if one assumes, for example, that your data observations have been contaminated by noise. This view is supported by Bi and Zhang (2005), which proposes the Total Support Vector Classifier (T-SVC). Specifically, as revealed by the RO formulation, we see that the T-SVC is equivalent to the non-convex EC-SVM with L_2 norm. In their paper, they show that by using an alternating minimization strategy to solve this problem, a special case of the non-convex EC-SVM performs better than the C-SVM if data have been contaminated by noise. Second, as we will discuss in the next section, the RO formulation suggests that the simple alternating minimization algorithm for finding the local minimum is equivalent to DCA in a certain sense.

10. Practical Considerations

In this section, we consider some practical implications of the theoretical insights connecting soft margin SVM's and bPOE minimization. We first discuss the most obvious use, which is for the selection of the threshold parameter. We show that the new interpretation of the C tradeoff parameter suggests a range that should be used to select the C parameter. This helps to improve grid selection for cross validation so that one can partially avoid solving the EC-SVM for values of C that yield trivial or redundant optimal solutions. Furthermore, using the robust interpretation of the EC-SVM, we show that there may be situations where prior knowledge about data uncertainty suggests, at-best, a fixed C or, at-worst, much tighter bounds for the range of interesting C values for cross-validation.

Another practical consideration is the optimization of the EC-SVM for the convex and non-convex case. For the convex case, we discuss alternative formulations that can be efficiently solved by standard convex optimization tools. With some of these approaches not well suited for large scale problems, we briefly mention recent work on subgradient methods that can be efficiently applied to the EC-SVM. For the non-convex case, we suggest a simple alternating minimization algorithm and show that this procedure has appealing connections with DCA, the Difference of Convex Functions Algorithm; see Tao and An (1997) or Dinh and Le Thi (2014).

Finally, with most learning applications, the true distribution of (X, y) is unknown and thus we must deal with only a finite sample. For example, throughout this paper, we assumed that we had N observations. Thus, in this case, we work with empirical estimates of bPOE, the quantile, the superquantile, and misclassification rate of the true distribution. A practical question, therefore, is regarding generalization. Given an optimal classifier by solving the EC-SVM (or equivalently the C-SVM), what can we say about the true rate of misclassification, $P(-y(w^T X + b) > 0)$, or the true value of bPOE, $\bar{p}_z(-y(w^T X + b))$? With regard to misclassification rate, we utilize results from Takeda and Sugiyama (2008); Schölkopf et al. (2000) to show that generalization bounds can be posed in terms of empirical estimates of bPOE, the quantile, and the superquantile, i.e. the information provided by Property 5. With regard to bPOE, we apply the stability arguments of Bousquet and Elisseeff (2002) to provide tight generalization bounds on the true bPOE at any non-positive threshold and show that the SVM parameter, C, plays an important role in these bounds.

10.1 Parameter Selection

10.1.1 Bounds for the Threshold

With the EC-SVM having a parameter interpretable as the bPOE threshold, we use this to select bounds on the admissible range of the parameter C. From (9), we see that for a random variable Z, the only interesting values of the parameter are $-C = z \in [E[Z], \sup Z]$. Finding this range, or a range containing it, can assist in performing grid search for model selection where we are trying to find the best choice of C without unnecessary iterations over meaningless or redundant parameter values (e.g. different parameter values all yielding bPOE equal to 1). First, we consider the EC-SVM with unbiased linear classifier, where $b = 0$. Then, using this result and some geometric intuition, we are able to formulate a similar bound for biased classifiers, where b is a free variable.

Since the EC-SVM is minimizing bPOE of $\frac{-y(w^TX+b)}{\|w\|}$, we know from (9) that the threshold parameter C achieves varying solutions only for

$$-C \in \left[\inf_{w,b} E\left[\frac{-y(w^TX+b)}{\|w\|} \right], \inf_{w,b} \left\{ \sup_{X,y} \left(\frac{-y(w^TX+b)}{\|w\|} \right) \right\} \right] .$$

For the case where $b = 0$, we can use simple algebraic arguments to find data dependent upper and lower bounds for this range. First, we have the lower bound,

$$-C \geq \min_w \frac{1}{N} \sum_i \frac{-y_i(w^TX_i)}{\|w\|} \geq -\max_{i,w} \frac{w^TX_i}{\|w\|} = -\max_i \|X_i\|^* .$$

Next, we find an upper bound for this range.

$$-C \leq \min_w \max_i \frac{-y_i(w^TX_i)}{\|w\|} \leq \max_{i,w} \frac{w^TX_i}{\|w\|} = \max_i \|X_i\|^* .$$

Therefore, when the EC-SVM is formulated with unbiased hyperplane w^TX, we only need to do a grid search for $-C \in [-\max_i \|X_i\|^*, \max_i \|X_i\|^*]$. Now, consider the case of the biased hyperplane. To find a lower bound of this range, first note, importantly, that by having a biased hyperplane (i.e. by including the b term), we have that the lower bound can be trivially unbounded with $\inf_{w,b} E\left[\frac{-y(w^TX+b)}{\|w\|}\right] = -\infty$. But this is unreasonable since this hyperplane trivially classifies all points as belonging to the same class. We know that any reasonable hyperplane should at least intersect the norm ball supporting the data set, i.e. the norm ball with radius $\max_i \|X_i\|^*$. Therefore, any non-trivial hyperplane should have an expected margin error larger than the margin error of the best classified point. Because the hyperplane at least intersects the norm ball at one point, the best classified point will have margin error no smaller than $-2\max_i \|X_i\|^*$. Therefore, a reasonable lower bound for $-C$ is,

$$-C \geq -2\max_i \|X_i\|^* .$$

For the upper bound, the exact same logic applies but w.r.t. the worst classified point for any hyperplane intersecting the norm ball containing all data points. Therefore, a reasonable upper bound for $-C$ is,

$$-C \leq 2\max_i \|X_i\|^* .$$

Thus, when the EC-SVM is formulated with biased hyperplane $w^T X + b$, we only need to do a grid search for $-C \in [-2\max_i \|X_i\|^*, 2\max_i \|X_i\|^*]$.

10.1.2 Using the Robust Interpretation

For specific situations, one can actually use the RO representation of the EC-SVM to select the proper parameter value. Since the parameter does not change value, along with the optimal (w, b), when moving from formulation (19) to (28) or (19) to (30), we can use the RO interpretation, exactly, to select parameter values for (19).

A situation where this would be ideal is if each data point X_i is given by some measurement and we know the maximum amount by which this measurement can deviate from the true value. Consider the following two dimensional example, where we use some equipment to measure blood pressure $X^{(1)}$ and weight $X^{(2)}$ where $X = [X^{(1)}, X^{(2)}] \in \mathbb{R}^2$. Assume that we know that the equipment to measure blood pressure is off by at most ±2 and that the equipment to measure weight is off by at most ±2. In this case, for any specific measurement X_i, we know that a representative uncertainty set would be $\|\delta_i\|_\infty \leq 2$. This immediately implies that $C = 2$ is desirable if one chooses to be *pessimistic* about the uncertainty or that $C = -2$ is desirable if one chooses to be *optimistic* about the uncertainty. Furthermore, this implies that one should use the EC-SVM formulated with the L_1 norm $\|w\|_1$ since its dual norm is the L_∞ norm.

Furthermore, in less idealistic situations when knowledge of the deviation of measurement is not so exact, the same intuition can be used to, at least, select a range of C values for cross validation that is smaller than the interval $[-2\max_i \|X_i\|^*, 2\max_i \|X_i\|^*]$ provided in the previous section.

10.2 Optimization of the EC-SVM

10.2.1 Convex Case

The primal and dual EC-SVM formulated with any general norm and $C \geq 0$ are convex optimization problems and, thus, there are many options for efficient optimization. Essentially, since from the optimization viewpoint the difference between C-SVM and EC-SVM is minor, the complexity of convex EC-SVM is exactly the same as that of C-SVM. For a general norm, due to convexity, gradient methods may be successfully implemented. For particular norms, we may reformulate the problem as a specific type of convex optimization problem to target available solvers. Below, we detail some problem reformulations into Quadratically Constrained Quadratic Programs (QCQP), Second-Order Cone Programs (SOCP), and Linear Programs (LP). If we assume that the norm is the euclidean norm, the EC-SVM with $C \geq 0$ can be written as the convex QCQP,

$$\begin{aligned} \min_{w,b,\xi,t} \quad & \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i + b) + Ct + 1, \quad \forall i \in \{1,\dots,N\}. \\ & \xi \geq 0 \\ & t^2 \geq w^T w. \end{aligned} \tag{31}$$

Note that if we simply write $t \geq \|w\|_2$, it is a convex SOCP. If the norm is LP representable (e.g. the L_1 or L_∞ norm), the problem can be reformulated as an LP. For example, with the L_1 norm, the EC-SVM with $C \geq 0$ can be written as,

$$\begin{aligned}
\min_{w,\hat{w},b,\xi,t} \quad & \sum_{i=1}^{N} \xi_i \\
s.t. \quad & \xi_i \geq -y_i(w^T X_i + b) + Ct + 1, \quad \forall i \in \{1, \dots, N\}. \\
& \xi \geq 0 \\
& t \geq \sum_{j=1}^{n} \hat{w}_j \\
& \hat{w}_j \geq w_j \\
& \hat{w}_j \geq -w_j.
\end{aligned} \tag{32}$$

Note that the convex dual EC-SVM can also be written as a QCQP, SOCP, or LP in a similar manner depending on the choice of norm.

For large-scale problems, though, these reformulations may suffer. For the C-SVM, this issue has garnered much attention. Although we leave in-depth exploration to future work, we note that many techniques used to solve large scale SVM's are directly applicable to the convex EC-SVM. Note that the primal EC-SVM with $C \geq 0$ can be formulated as the following convex, non-smooth optimization problem.

$$\min_{w,b} \sum_{i=1}^{N} [-y_i(w^T X_i + b) + C\|w\| + 1]^+ \tag{33}$$

This formulation lends itself to subgradient methods, particularly stochastic variants which have been shown to work well for large-scale SVMs. In fact, methods mirroring Pegasos Shalev-Shwartz et al. (2011), a solver using a subgradient method which finds an ϵ-accurate solution, with high probability, for the primal C-SVM in runtime $O(\frac{1}{C\epsilon})$, have been successfully applied to large-scale robust SVM's taking a similar form to (33) in Wang et al. (2016).[4] Note, also, that the EC-SVM can be formulated as a decomposable consensus problem which can be solved in a distributed way via the Alternating Direction Method of Multipliers; see Chapter 7 and 8 of Boyd et al. (2011).

10.2.2 Non-Convex Case

For the EC-SVM with $C < 0$, solving for a global optimum is much more difficult. We can, though, solve for a local minimum by using the optimistic RO formulation (30) and implementing the simple, seemingly heuristic, approach of alternating minimization. While this approach was shown to be effective for the T-SVC special case of the EC-SVM presented in Bi and Zhang (2005), we show here that its effectiveness may be explained, on the surface at least, by its connection with DCA. Specifically, we have the following alternating algorithm, where at each step one solves an LP followed by a convex optimization problem:

[4] Ignoring some details for simplicity, the robust SVM of similar form is formulated as $\min_{w,b} \sum_i [-y_i(w^T X_i + b) + C\|\Sigma^{\frac{1}{2}} w\| + 1]^+$, where Σ is a particular covariance matrix.

- **Initialize:** Minimize (30) w.r.t. (w, b), fixing $\delta_i = 0$ for all $i = 1, ..., N$. Denote optimal point as (w^*, b^*).
- **Step 1:** Let $\hat{\delta}_i = \underset{\|\delta_i\|^* \leq -C}{\text{argmax}}\ w^{*T}\delta_i$ for all $i = 1, ..., N$.
- **Step 2:** Minimize (30) w.r.t. (w, b), fixing $\delta_i = \hat{\delta}_i$ for all $i = 1, ..., N$. Denote the optimal point as (w^*, b^*).
- **Step 3:** If iteration limit reached or convergence criteria is satisfied, STOP. Else, return to Step 1.

This algorithm, at worst, seems like a simple heuristic approach to a non-convex problem with intuitive appeal. In Step 1, we start with the data $X + \delta$, which is an optimistic version of the original data X if $\delta \neq 0$. We then find a classifier (w, b) that minimizes bPOE at threshold zero on the optimistic version of the data set. Then, in Step 2, we look for a new, optimistic view of the data that is optimistic w.r.t. the new classifier found in Step 1.

In addition to its intuitive appeal, however, this alternating minimization has theoretical justification. We can show that this alternating minimization algorithm is essentially implementing DCA, a commonly used algorithm for solving DC optimization problems. In short, a DC function h, is such that $h(v) = h_1(v) - h_2(v)$ where h_1, h_2 are both convex. To find a local minimum for the problem $\min_v h(v)$, DCA performs the following steps[5] :

- **Initialize:** $\hat{v} = \underset{v}{\text{argmin}}\ h_1(v)$.
- **Step 1:** Set δ equal to a subgradient of h_2 at $\hat{v}$: $\delta \in \partial h_2(\hat{v})$.
- **Step 2:** Set $\hat{v} = \underset{v}{\text{argmin}}\ h_1(v) - v^T\delta$.
- **Step 3:** If iteration limit reached or convergence criteria is satisfied, STOP. Else, return to Step 1.

Thus, DCA can be seen as first linearizing the concave term at the current solution via a subgradient and then solving a convex subproblem which replaces the concave term with the linear approximation.

To see how the alternating minimization procedure is related to DCA, first notice that when $C < 0$, the function $-y_i(w^T X_i + b) + C\|w\| + 1$ is DC; the sum of the convex function $-y_i(w^T X_i+b)+1$ and the concave function $C\|w\|$. Therefore, $\sum_i[-y_i(w^T X_i+b)+C\|w\|+1]^+$ is also DC.[6] Second, notice that $-C\|w\|$ can be viewed as the support function of the convex set $S = \{\delta | \|\delta\|^* \leq -C\}$, meaning that $-C\|w\| = \max_{\delta \in S} w^T\delta$. By the properties of support functions[7] we then know that if $\delta \in \underset{\|\delta\|^* \leq -C}{\text{argmax}}\ w^T\delta$, then $\delta \in \partial(-C\|w\|)$.

Therefore, we see that we initialize the alternating minimization algorithm by setting the concave part to zero and minimizing the convex part. Then, in Step 1, because of the

[5] Note that for brevity, we present a very simplified view of DCA and ignore the underlying duality results which produce this algorithm.

[6] This follows from the property that the maximum of DC functions is DC and that the sum of DC functions is DC.

[7] See e.g. Chapter 13 of Rockafellar (2015).

properties of the support function, we are equivalently solving for $\hat{\delta}_i \in \partial(-C\|w\|)$. Finally, in Step 2, we linearize the concave part using this subgradient and solve the resulting convex problem. Thus, not only is the alternating minimization algorithm simple, but it is also theoretically appealing with its relation to DCA.

10.3 Generalization of POE and bPOE for SVM Classifiers

In machine learning, particularly in the context of empirical risk minimization, statistical learning theory is a popular tool for analyzing the predictive ability of a learning algorithm $A : S \to \mathcal{F}$ which maps a sample S onto a function $A_S \in \mathcal{F}$, with $\mathcal{F}$ being some class of functions such that the score $A_S(X) \in \mathbb{R}$ is used to predict y given X. For example, in classification, we can consider the prediction of y given X to be $sign\{A_S(X)\}$. Statistical learning theory addresses the fact that the true distribution of (X, y) is usually unknown and we are only given access to a finite sample $S = \{(X_1, y_1), ..., (X_N, y_N)\}$. Thus, if we have, for example, a real valued random loss $V(A_S, (X, y))$ for the output A_S of the learning algorithm trained on S, we need to estimate the true risk

$$\mathcal{R}_{true}(A, S) = E_{X,y}[V(A_S, (X, y))] ,$$

given only the empirical estimate,

$$\mathcal{R}_{emp}(A, S) = \frac{1}{N} \sum_i V(A_S, (X_i, y_i)) .$$

The risk can be viewed as a type of performance metric we would like to minimize, like bPOE or misclassification rate. Given a training algorithm and a training set S, statistical learning theory helps to determine the expected value of the difference $|\mathcal{R}_{true}(A, S) - \mathcal{R}_{emp}(A, S)|$.

In the context of binary classification, we consider two different types of risk (or performance metrics) for the EC-SVM (or equivalently C-SVM) learning algorithm. We first consider the true misclassification rate, which is the typical object of study for classification algorithms in statistical learning theory. Drawing on results from Takeda and Sugiyama (2008), we show that, given a solution to the EC-SVM, the true misclassification rate can be upper bounded by a function of the empirical estimates of bPOE, the superquantile, and the quantile of the random loss $\left(\frac{-y(w^T X+b)}{\|w\|}\right)$. Furthermore, given Property 5, we find that we can state these bounds in terms of the value of the EC-SVM objective function, C, and $\frac{1}{\|w\|}$.

Next, instead of misclassification rate, we consider generalization bounds for the true bPOE of the random loss $\left(\frac{-yw^T X}{\|w\|}\right)$. Using the stability arguments of Bousquet and Elisseeff (2002), given a solution to the EC-SVM or C-SVM, we provide tight upper bounds for the true bPOE at any non-positive threshold. In general, for binary classification, we find that using these stability arguments to bound bPOE is more intuitive than the use of the clipped loss function from Bousquet and Elisseeff (2002). Note that we only consider the unbiased case with $b = 0$, as it is unclear whether this bound holds for the biased case. Furthermore, in Section 10.3.2 we consider only the L_2 norm and in Section 10.3.3 we only consider norms possessing a variant of strong convexity.

In the following sections, given a sample $S = \{(X_1, y_1), ..., (X_N, y_N)\}$, we consider $L(w, b, S)$ to be a discretely distributed random loss with equally probable realizations $\{-y_1(w^T X_1 + b), ..., -y_N(w^T X_N + b)\}$. Therefore, we have that

$$\bar{p}_{-C}\left(\frac{L(w,b,S)}{\|w\|}\right), \bar{q}_\alpha\left(\frac{L(w,b,S)}{\|w\|}\right), \text{ and } q_\alpha\left(\frac{L(w,b,S)}{\|w\|}\right)$$

are empirical estimates of the *true* bPOE, superquantile, and quantile

$$\bar{p}_{-C}\left(\frac{-y(w^T X + b)}{\|w\|}\right), \bar{q}_\alpha\left(\frac{-y(w^T X + b)}{\|w\|}\right), q_\alpha\left(\frac{-y(w^T X + b)}{\|w\|}\right) .$$

10.3.1 Generalization of POE

Takeda and Sugiyama (2008) provide bounds on the true misclassification rate of the Eν-SVM classifier. Given the equivalence between the Eν-SVM and EC-SVM provided in Section 7, we can apply these results directly to the EC-SVM. Although these results are pulled directly from Takeda and Sugiyama (2008), hence we do not provide the proofs, we find that we can pose their bounds on the true misclassification rate strictly in terms of the empirical estimates of bPOE, the superquantile, and the quantile of the random loss $\left(\frac{-y(w^T X+b)}{\|w\|_2}\right)$. Furthermore, given Property 5, we find that we can state these bounds in terms of the value of the EC-SVM objective function, C, and $\frac{1}{\|w\|_2}$. Note also that we are able to provide bounds for the non-convex case where $C < 0$.

For the following theorems, let $P_{true} = P\left(-y(w^T X + b) > 0\right)$, i.e. the true misclassification rate for the classifier (w, b), and let

$$G(\tau) = \sqrt{\frac{2}{N}\left(\frac{4k^2(R^2+1)^2}{\tau^2}\log_2(2N) - 1 + \ln\frac{2}{\rho}\right)} .$$

Theorem 3 *Suppose that (w, b) is an optimal solution to the EC-SVM trained on sample S, that the optimal objective value divided by sample size N equals $1 - \alpha$, and that either $C > 0$ or $C < 0$ and $-C - \frac{1}{\|w\|_2} < 0$. If the support $\mathcal{X}$ of X is contained in a ball of radius R, then there exists a positive constant k such that the following bound holds with probability at least $1 - \rho$:*

$$P_{true} \leq \bar{p}_{-C}\left(\frac{L(w,b,S)}{\|w\|_2}\right) + G\left(q_\alpha\left(\frac{L(w,b,S)}{\|w\|_2}\right)\right) .$$

Furthermore, given Property 5, this can be written simply as:

$$P_{true} \leq 1 - \alpha + G\left(-C - \frac{1}{\|w\|_2}\right) .$$

Theorem 4 *Make the same assumptions as in Theorem 3, but suppose that $C < 0$ and $-C - \frac{1}{\|w\|_2} > 0$. Then there exists a positive constant k such that the following bound holds with probability at least $1 - \rho$:*

$$P_{true} \geq \bar{p}_{-C}\left(\frac{L(w,b,S)}{\|w\|_2}\right) - G\left(q_\alpha\left(\frac{L(w,b,S)}{\|w\|_2}\right)\right) .$$

Furthermore, given Property 5, this can be written simply as:

$$P_{true} \geq 1 - \alpha - G\left(-C - \frac{1}{\|w\|_2}\right) .$$

Given these theorems, we see that the EC-SVM is directly minimizing the the lower (or upper) bound on P_{true}, while simultaneously trading off w.r.t. the choice in parameter C.

10.3.2 Generalization of Buffered Probability

Here, we use the stability arguments of Bousquet and Elisseeff (2002) to provide generalization bounds on bPOE for optimal solutions of the C-SVM and convex EC-SVM. We provide tight bounds for bPOE of the normalized loss distribution of the classifier given by an SVM trained on a finite sample. This shows that while the C-SVM and EC-SVM directly minimize bPOE, this statistic does, in fact, generalize to unseen samples.

For brevity, we avoid a full introduction of uniform stability and the associated generalization theorems and refer readers to Bousquet and Elisseeff (2002). Thus, we confine most details to the proof. Theorem 5 considers the C-SVM, but with squared norm, and shows that we can provide tight bounds on the true bPOE of the loss $\left(\frac{-yw^TX}{\|w\|}\right)$ at any non-positive threshold. We present the C-SVM with squared norm in the main theorem, as opposed to the EC-SVM, because the bound expression is much more straightforward. Corollary 1 shows how this result applies to the case of the EC-SVM with L_2 norm.

In general, both bound expressions provide the same intuitive result. The uncertainty regarding the estimate of bPOE will grow as the bPOE threshold considered in the SVM minimization shrinks. As the SVM parameter grows larger, the threshold for the equivalent bPOE minimization problem decreases. Thus, we are considering a growing portion of the tail of the loss distribution in our minimization. This means that the SVM objective function will get larger, but the uncertainty will grow smaller since a larger portion of the tail will be easier to estimate than a smaller portion (it contains more samples). The opposite holds true as the SVM parameter shrinks, getting closer to zero. This implies that we are considering a shrinking portion of the tail of the loss distribution in our minimization. Thus, the SVM objective function will grow smaller, but the uncertainty will grow larger since a smaller portion of the tail will be more difficult to empirically estimate (it contains less samples).

We note that we assume that the classifier is unbiased with $b = 0$ and that the squared norm is differentiable and m-strongly convex w.r.t. itself.[8] For example, the squared L_2 norm is strongly convex w.r.t. itself with $m = 2$. A bias term can be included by concatenating a feature equal to one onto the data, but this would then be included in the regularization norm. It is unclear whether these bounds hold for the truly unbiased case. Although we do not present it, we also point out that this can easily be generalized to consider functions in a Reproducing Kernel Hilbert Space such that the kernel function $K(\cdot,\cdot)$ is bounded with $\sup_{X\in\mathcal{X}} K(X,X) \leq R$, as is done in Bousquet and Elisseeff (2002). For the following proposition, recall that $\bar{p}_{-C}\left(\frac{-yw^TX}{\|w\|}\right)$ denotes the true bPOE while $\bar{p}_{-C}\left(\frac{L(w,S)}{\|w\|}\right)$ denotes empirical bPOE.

[8] A differentiable function f is m-strongly convex w.r.t. the norm $\|\cdot\|$ if for any w_1, w_2 we have that $\frac{m}{2}\|w_1 - w_2\|^2 \leq f(w_1) - f(w_2) - \langle \nabla f(w_2), w_1 - w_2\rangle$.

Theorem 5 *Assume that the squared norm $\| \cdot \|^2$ is differentiable and m-strongly convex w.r.t. itself and that $\sup_{x \in \mathcal{X}} \|x\|^* \leq R$, meaning that the support $\mathcal{X}$ of X is contained in a $\| \cdot \|^*$-norm ball of radius R. Assume we solve the C-SVM with squared norm on training set $S = \{(X_1, y_1), ..., (X_N, y_N)\}$, yielding*

$$w_S \in \underset{w}{\mathit{argmin}}\ \frac{1}{N}\sum_i [-y_i w^T X_i + 1]^+ + \lambda\|w\|^2 \ ,$$

where $\lambda > 0$. Then for any $a, C \geq 0$, with probability $1 - \rho$,

$$\bar{p}_{-C}\left(\frac{-y w_S^T X}{\|w_S\|}\right) \leq \frac{1}{N}\sum_i [-y_i a w_S^T X_i + C\|a w_S\| + 1]^+ + \frac{U_{\frac{1}{2}}(a, \lambda, C)\sqrt{\ln \frac{1}{\rho}}}{\sqrt{N}} + \frac{U_1(a, \lambda, C)}{N} \ ,$$

where $U_{\frac{1}{2}}(a, \lambda, C) = \frac{1}{\sqrt{2}} + \frac{a(R+C)}{\sqrt{2\lambda}}\left(1 + \frac{4R}{m\sqrt{\lambda}}\right)$, $U_1(a, \lambda, C) = \frac{2aR(R+C)}{m\lambda}$. In particular,

$$\bar{p}_{-C}\left(\frac{-y w_S^T X}{\|w_S\|}\right) \leq \bar{p}_{-C}\left(\frac{L(w_S, S)}{\|w_S\|}\right) + \frac{U_{\frac{1}{2}}(\bar{a}_S, \lambda, C)\sqrt{\ln \frac{1}{\rho}}}{\sqrt{N}} + \frac{U_1(\bar{a}_S, \lambda, C)}{N},$$

with probability $1 - \rho$, when $\bar{a}_S \in \underset{a \geq 0}{\mathit{argmin}}\ \frac{1}{N}\sum_i [-y_i a w_S^T X_i + C\|a w_S\| + 1]^+$.

Proof Let $V(w, (X, y), a, C) = [-y a w^T X + C\|a w\| + 1]^+$. First, note that, for any $C \geq 0$,

$$\begin{aligned}
\bar{p}_{-C}\left(\frac{-y w_S^T X}{\|w_S\|}\right) &= \min_{a \geq 0} E[a\left(\frac{-y w_S^T X}{\|w_S\|} + C\right) + 1]^+ \\
&= \min_{a \geq 0} E[\frac{a}{\|w_S\|}\left(-y w_S^T X + C\|w_S\|\right) + 1]^+ \\
&= \min_{a \geq 0} E[-y a w_S^T X + C\|a w_S\| + 1]^+ \\
&\leq E[V(w_S, (X, y), a, C)] \ , \forall\, a \geq 0 \ .
\end{aligned}$$

Given any $a \geq 0$, we can upper bound $E[V(w_S, (X, y), a, C)]$ by using stability arguments of Bousquet and Elisseeff (2002).

We first prove that the C-SVM with squared norm is uniformly stable, as defined in Bousquet and Elisseeff (2002), w.r.t. loss function $V(w, (X, y), a, C)$, showing that there exists θ such that for any sample S,

$$\sup_{X, y} |V(w_S, (X, y), a, C) - V(w_{S^j}, (X, y), a, C)| \leq \theta$$

where

$$w_S \in \underset{w}{\operatorname{argmin}}\ \frac{1}{N}\sum_i [-y_i w^T X_i + 1]^+ + \lambda\|w\|^2$$

$$w_{S^j} \in \underset{w}{\operatorname{argmin}}\ \frac{1}{N}\sum_{i \neq j} [-y_i w^T X_i + 1]^+ + \lambda\|w\|^2 \ .$$

First, note that for any $w_1, w_2 \in \mathbb{R}^n$, $X, y \in \mathcal{X} \times \{-1, 1\}$ we have that

$$\begin{aligned}
|V(w_1,(X,y),a,C) - V(w_2,(X,y),a,C)| &= |[-y(aw_1^T X) + C\|aw_1\| + 1]^+ \\
&\qquad - [-y(aw_2^T X) + C\|aw_2\| + 1]^+| \\
&\leq | - y(aw_1^T X) + C\|aw_1\| + 1 \\
&\qquad - \big(-y(aw_2^T X) + C\|aw_2\| + 1\big) | \\
&= \left|-ya(w_1 - w_2)^T X + Ca(\|w_1\| - \|w_2\|)\right| \\
&\leq \left|-ya(w_1 - w_2)^T X\right| + Ca\left|\|w_1\| - \|w_2\|\right| \\
&\leq \left|-ya(w_1 - w_2)^T X\right| + Ca\|w_1 - w_2\| \\
&\leq a\left|(w_1 - w_2)^T X\right| + Ca\|w_1 - w_2\| \\
&\leq \sup_{X \in \mathcal{X}} a\left|(w_1 - w_2)^T X\right| + Ca\|w_1 - w_2\| \\
&\leq \sup_{\|X\|^* \leq R, X \in \mathbb{R}^n} a\left|(w_1 - w_2)^T X\right| + Ca\|w_1 - w_2\| \\
&= Ra\|w_1 - w_2\| + Ca\|w_1 - w_2\| \\
&= a(R + C)\|w_1 - w_2\|
\end{aligned}$$

We show now that $\|w_S - w_{S^j}\| \leq \frac{R}{mN\lambda}$. Let $d_f, \bar{d}_f$ denote the Bregman divergence and the generalized Bregman divergence of a convex function f with subderivatives at any w denoted as $\nabla f(w) \in \partial f(w)$. (See Appendix F for definition and properties of divergence). Let $g(w) = \|w\|^2$, $h_S(w) = \frac{1}{N}\sum_i [-y_i w^T X_i + 1]^+$, and $h_{S^j}(w) = \frac{1}{N}\sum_{i \neq j}[-y_i w^T X_i + 1]^+$. First, since $g(w)$ is assumed to be m-strongly convex w.r.t. itself, by definition we have $\frac{m}{2}\|w_S - w_{S^j}\|^2 \leq d_g(w_S, w_{S^j})$. Next, we have that

$$\begin{aligned}
\lambda m\|w_S - w_{S^j}\|^2 &\leq \lambda\left(d_g(w_{S^j}, w_S) + d_g(w_S, w_{S^j})\right) \\
&= d_{\lambda g}(w_{S^j}, w_S) + d_{\lambda g}(w_S, w_{S^j}) \\
\text{(following from (44))} &= \bar{d}_{\lambda g}(w_{S^j}, \nabla\lambda g(w_S)) + \bar{d}_{\lambda g}(w_S, \nabla\lambda g(w_{S^j}))
\end{aligned}$$

Since generalized divergence is linear and non-negative (see Appendix F and (45)), we have for any $\nabla(h_S(w_S) + \lambda g(w_S)) \in \partial(h_S(w_S) + \lambda g(w_S))$ and any $\nabla(h_{S^j}(w_{S^j}) + \lambda g(w_{S^j})) \in \partial(h_{S^j}(w_{S^j}) + \lambda g(w_{S^j}))$, that

$$\begin{aligned}
\bar{d}_{\lambda g}(w_{S^j}, \nabla\lambda g(w_S)) + \bar{d}_{\lambda g}(w_S, \nabla\lambda g(w_{S^j})) &\leq \bar{d}_{(h_S + \lambda g)}(w_{S^j}, \nabla(h_S(w_S) + \lambda g(w_S))) \\
&\quad + \bar{d}_{(h_{S^j} + \lambda g)}(w_S, \nabla(h_{S^j}(w_{S^j}) + \lambda g(w_{S^j}))) \,.
\end{aligned}$$

Now, since $0 \in \partial(h_S(w_S) + \lambda g(w_S))$ and $0 \in \partial(h_{S^j}(w_{S^j}) + \lambda g(w_{S^j}))$, we have

$$\begin{aligned}
\bar{d}_{\lambda g}(w_{S^j}, \nabla\lambda g(w_S)) + \bar{d}_{\lambda g}(w_S, \nabla\lambda g(w_{S^j})) &\leq \bar{d}_{(h_S + \lambda g)}(w_{S^j}, 0) + \bar{d}_{(h_{S^j} + \lambda g)}(w_S, 0) \\
\text{(following from (46))} &= h_S(w_{S^j}) + \lambda g(w_{S^j}) - h_S(w_S) - \lambda g(w_S) \\
&\quad + h_{S^j}(w_S) + \lambda g(w_S) - h_{S^j}(w_{S^j}) - \lambda g(w_{S^j}) \\
&= h_S(w_{S^j}) - h_S(w_S) + h_{S^j}(w_S) - h_{S^j}(w_{S^j}) \\
&= \frac{1}{N}\left(V(w_{S^j}, (X_j, y_j), 1, 0) - V(w_S, (X_j, y_j), 1, 0)\right) \\
&\leq \frac{R\|w_{S^j} - w_S\|}{N}
\end{aligned}$$

This then implies that $\|w_S - w_{S^j}\|^2 \leq \|w_S - w_{S^j}\| \frac{R}{mN\lambda}$ which further implies that $\|w_S - w_{S^j}\| \leq \frac{R}{N\lambda m}$. Together with the previous result, we have that

$$|V(w_S, (X,y), a, C) - V(w_{S^j}, (X,y), a, C)| \leq a(R+C)\|w_S - w_{S^j}\| \leq \frac{aR(R+C)}{N\lambda m} \ .$$

Thus, the learning algorithm is uniformly θ-stable w.r.t. loss function V with $\theta = \frac{aR(R+C)}{N\lambda m}$.

To apply Theorem 12 of Bousquet and Elisseeff (2002), which gives us the final generalization bound, we need to show that $V(w_S, (X,y), a, C)$ is bounded. To see this, first notice that

$$\lambda\|w_S\|^2 \leq h_S(w_S) + \lambda\|w_S\|^2 \leq h_S(0) + \lambda\|0\|^2 = \frac{1}{N}\sum_i V(0, (X_i, y_i), 1, 0) = 1 \ .$$

Second, combining this with the first result of the proof yields,

$$\begin{aligned} V(w_S, (X,y), a, C) &\leq |V(w_S, (X,y), a, C) - V(0, (X,y), a, C)| + V(0, (X,y), a, C) \\ &\leq a(R+C)\|w_S - 0\| + 1 \\ &\leq \frac{a(R+C)}{\sqrt{\lambda}} + 1 \ . \end{aligned}$$

Thus, applying Theorem 12 of Bousquet and Elisseeff (2002), for any $a \geq 0$ yields the first generalization bound. The second generalization bound follows from the fact that if

$$\bar{a} \in \operatorname*{argmin}_{a \geq 0} \frac{1}{N}\sum_i [-y_i a w_S^T X_i + C\|a w_S\| + 1]^+ \ ,$$

then,

$$\frac{1}{N}\sum_i [-y_i \bar{a} w_S^T X_i + C\|\bar{a} w_S\| + 1]^+ = \bar{p}_{-C}\left(\frac{L(w_S, S)}{\|w_S\|}\right) \ .$$

■

Corollary 1 *Make the same assumptions as in Theorem 5, but solve the EC-SVM with $\|\cdot\| = \|\cdot\|_2$ and $\hat{C} > 0$, which yields,*

$$w_S \in \mathop{argmin}_{w} \frac{1}{N}\sum_i [-y_i(w^T X_i) + 1 + \hat{C}\|w\|_2]^+ \ .$$

Then, for any $a, C \geq 0$, with probability $1-\rho$,

$$\bar{p}_{-C}\left(\frac{-y w_S^T X}{\|w_S\|}\right) \leq \frac{1}{N}\sum_i [-y_i a\mu w_S^T X_i + C\|a\mu w_S\| + 1]^+ + \frac{U_{\frac{1}{2}}(a, \lambda_S, C)\sqrt{\ln\frac{1}{\rho}}}{\sqrt{N}} + \frac{U_1(a, \lambda_S, C)}{N},$$

where $\lambda_S = \frac{\hat{C}\sum_i \beta_i}{2\|w_S\|_2}$, $\mu = \frac{1}{1+\hat{C}\|w_S\|_2}$, *and* β_i *are the optimal dual variables of the EC-SVM. In particular, when* $\bar{a}_S \in \underset{a\geq 0}{\mathit{argmin}}\ \frac{1}{N}\sum_i[-y_i a\mu w_S^T X_i + C\|a\mu w_S\| + 1]^+$, *with probability* $1-\rho$,

$$\bar{p}_{-C}\left(\frac{-y w_S^T X}{\|w_S\|}\right) \leq \bar{p}_{-C}\left(\frac{L(w_S,S)}{\|w_S\|}\right) + \frac{U_{\frac{1}{2}}(\bar{a}_S,\lambda_S,C)\sqrt{\ln\frac{1}{\rho}}}{\sqrt{N}} + \frac{U_1(\bar{a}_S,\lambda_S,C)}{N}.$$

Proof From Theorem 2, we know that if w is optimal for the EC-SVM with $\hat{C} \geq 0$, then solving the EC-SVM is equivalent to solving the C-SVM with parameter $\hat{C}\sum_i \beta$ and will yield optimal solution μw_S. Furthermore, it is easy to see by looking at the KKT system (34) that if w is optimal for the C-SVM with parameter $\hat{C} \geq 0$, then solving the C-SVM is equivalent to solving the C-SVM with squared norm with parameter $\frac{\hat{C}}{2\|w\|_2}$ and will yield optimal solution w. Therefore, if w_S is optimal for the EC-SVM with $\hat{C} \geq 0$, then solving the EC-SVM is equivalent to solving the C-SVM with squared norm with parameter $\frac{\hat{C}\sum_i \beta_i}{2\|w_S\|_2}$ and will yield optimal solution μw. Thus, we can apply Theorem 5 with $\lambda = \frac{\hat{C}\sum_i \beta_i}{2\|w_S\|_2}$ and μw_S. ■

11. 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, 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 two useful interpretations of the EC-SVM. First, we showed that it can be considered to be a hard-margin optimization problem, showing that the choice of regularization norm implies a metric used to define the margin. Second, we showed that the convex and non-convex case can be interpreted as Robust Optimization problems, with convexity implying pessimistic views of data uncertainty and non-convexity implying optimistic views of data uncertainty.

We also address some practical implications of these theoretical insights. We show that the new interpretations imply that the interesting values of C lie in a limited range. We also suggest methods for solving the convex and non-convex case of the EC-SVM, using the robust interpretation in particular for solving the non-convex case in an efficient and theoretically justifiable way. Finally, we also show that we can provide generalization bounds for both misclassification rate and bPOE. We provide tight bounds for bPOE of the loss distribution of the classifier given by the C-SVM and EC-SVM trained on a finite sample. This shows that while the C-SVM and EC-SVM minimize bPOE, this statistic does, in fact, generalize to unseen samples.

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 little 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. Junya Gotoh, 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.

Appendix 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 2 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 $\min\limits_{w \in \mathbb{R}^n} \quad \bar{p}_z(f(w,X)) = 1 - \alpha^*$ at (w^*, γ^*), then for any $\bar{z} = az$, $a \geq 1$ we have that $\min\limits_{w \in \mathbb{R}^n} \quad \bar{p}_{\bar{z}}(f(w,X)) = 1 - \alpha^*$ at $(aw^*, a\gamma^*)$. ■

Corollary 2 *Given Proposition 4, 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 2 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 2 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. ■

Appendix 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):

$$\xi^* \geq 0, \tag{34a}$$

$$\beta^* \geq 0, \tag{34b}$$

$$\alpha^* \geq 0, \tag{34c}$$

$$-\alpha^* - \beta^* + 1 = 0, \tag{34d}$$

$$\xi_i^* \geq -y_i({w^*}^T x_i + b^*) + 1, \tag{34e}$$

$$0 \in -\hat{C}\partial\|w^*\| + \sum_{i=1}^{N} y_i \beta_i^* x_i, \tag{34f}$$

$$\sum_{i=1}^{N} -y_i \beta_i^* = 0, \tag{34g}$$

$$\alpha_i^* \xi_i^* = 0 = (1 - \beta_i^*)\xi_i^*, \tag{34h}$$

$$\beta_i^* \left[-y_i({w^*}^T x_i + b^*) + 1 - \xi_i^* \right] = 0. \tag{34i}$$

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:

$$\xi \geq 0, \tag{35a}$$

$$\beta \geq 0, \tag{35b}$$

$$\alpha \geq 0, \tag{35c}$$

$$-\alpha - \beta + 1 = 0, \tag{35d}$$

$$\xi_i \geq -y_i(w^T x_i + b) + 1 + C\|w\|, \tag{35e}$$

$$0 \in -C\partial\|w\| \sum_{i=1}^{N} \beta_i + \sum_{i=1}^{N} y_i \beta_i x_i, \tag{35f}$$

$$\sum_{i=1}^{N} -y_i \beta_i = 0, \tag{35g}$$

$$\alpha_i \xi_i = 0 = (1 - \beta_i)\xi_i, \tag{35h}$$

$$\beta_i \left[-y_i(w^T x_i + b) + 1 + C\|w\| - \xi_i \right] = 0. \tag{35i}$$

When now show 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 (35).

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 (34e).

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 $C\partial\|w\| \sum_{i=1}^N \beta_i = \hat{C}\partial\|\mu w^*\| = \hat{C}\partial\|w^*\|$, $\sum_{i=1}^N y_i\beta_i x_i = \sum_{i=1}^N y_i\beta_i^* x_i$, and (34f).

7. $\sum_{i=1}^N -y_i\beta_i = 0$: True, since $\beta = \beta^*$ and (34g).

8. $\alpha_i\xi_i = 0$: True, since $\alpha = \alpha^*$, $\xi = \mu\xi^*$, $\mu > 0$, and (34h).

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 (35h) $\implies$ $\beta_i\xi_i = \xi_i$. This gives us $\xi_i = \mu\xi^* = -\mu\beta_i^* y_i(w^{*T} x_i + b^*) + \beta_i^* + \beta_i^* C\|\mu w^*\| = \beta_i^* + \beta_i^* \frac{\hat{C}\mu}{\sum_{i=1}^N \beta_i^*}\|w^*\| + \mu\left[\xi_i^* - \beta_i^*\right]$ where $-\mu\beta_i^* y_i(w^{*T} x_i + b^*) = \mu\left[\xi_i^* - \beta_i^*\right]$ is implied by (35i). Furthermore, notice that this last equality is true $\iff$ $\beta_i^*\left[1 + \frac{\hat{C}\mu}{\sum_{i=1}^N \beta_i^*}\|w^*\| - \mu\right] = 0 \iff 1 + \frac{\hat{C}\|w^*\|}{\sum_{i=1}^N \beta_i^* - \hat{C}\|w^*\|} - \frac{\sum_{i=1}^N \beta_i^*}{\sum_{i=1}^N \beta_i^* - \hat{C}\|w^*\|} = 0$. Clearly, the last equality is true.

Now we show that for non-linearly separable data sets, $\mu > 0$. We use the KKT system (34) 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 (34d,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 (34d,g).

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^N \beta_i + \left(\inf_{w,b} \hat{C}\|w\| + \sum_{i=1}^N \beta_i\left[-y_i(w^T x_i + b)\right]\right) \\ s.t. \quad & 0 \leq \beta \leq 1. \\ & \sum_{i=1}^N y_i\beta_i = 0. \end{aligned} \tag{36}$$

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

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

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 $C \geq 0$ 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 (19):

$$\xi^* \geq 0, \tag{37a}$$
$$\beta^* \geq 0, \tag{37b}$$
$$\alpha^* \geq 0, \tag{37c}$$
$$-\alpha^* - \beta^* + 1 = 0, \tag{37d}$$
$$\xi_i^* \geq -y_i({w^*}^T x_i + b^*) + 1 + C\|w^*\|, \tag{37e}$$
$$0 \in -C\partial\|w^*\| \sum_{i=1}^{N} \beta_i^* + \sum_{i=1}^{N} y_i \beta_i^* x_i, \tag{37f}$$
$$\sum_{i=1}^{N} -y_i \beta_i^* = 0, \tag{37g}$$
$$\alpha_i^* \xi_i^* = 0 = (1 - \beta_i^*)\xi_i^*, \tag{37h}$$
$$\beta_i^* \left[-y_i({w^*}^T x_i + b^*) + 1 + C\|w^*\| - \xi_i^*\right] = 0. \tag{37i}$$

Now, we show that with $\mu = \frac{1}{1+C\|w^*\|} > 0$, the variables $(\mu w^*, \mu b^*, \mu \xi^*, \alpha^*, \beta^*)$ 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:

$$\xi \geq 0, \tag{38a}$$

$$\beta \geq 0, \tag{38b}$$

$$\alpha \geq 0, \tag{38c}$$

$$-\alpha - \beta + 1 = 0, \tag{38d}$$

$$\xi_i \geq -y_i(w^T x_i + b) + 1, \tag{38e}$$

$$0 \in -\hat{C}\partial\|w\| + \sum_{i=1}^{N} y_i \beta_i x_i, \tag{38f}$$

$$\sum_{i=1}^{N} -y_i \beta_i = 0, \tag{38g}$$

$$\alpha_i \xi_i = 0 = (1 - \beta_i)\xi_i, \tag{38h}$$

$$\beta_i \left[-y_i(w^T x_i + b) + 1 - \xi_i \right] = 0. \tag{38i}$$

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 (38):

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 (37e).

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 (37f).

7. $\sum_{i=1}^{N} -y_i \beta_i = 0 :$ True, since $\beta = \beta^*$ and (37g).

8. $\alpha_i \xi_i = 0 :$ True, since $\alpha = \alpha^*$, $\xi = \mu \xi^*$, $\mu > 0$, and (37h).

9. $\beta_i \left[-y_i(w^T x_i + b) + 1 - \xi_i \right] = \beta_i^* \left[-\mu y_i(w^{*T} x_i + b^*) + 1 - \mu \xi_i^* \right] =$ $\beta_i^* + \frac{\beta_i^* \left(-y_i(w^{*T} x_i + b^*) - \xi_i^* \right)}{1 + C\|w^*\|} = 0 \iff \beta_i^* \left[-y_i(w^{*T} x_i + b^*) + 1 + C\|w^*\| - \xi_i^* \right] = 0,$ which is True, following from (37i).

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

Appendix C. Derivation of EC-SVM Dual Formulation

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

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

Appendix D. Derivation of C-SVM Dual Formulation

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

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

Appendix E. Derivation of Formula (21)

To see how we derive (21), we begin with formula (14) with $f(w,X) = -y(w^TX+b)$ and $z = -C$:

$$
\begin{aligned}
\min_{\gamma<-C,w,b} \quad & \sum_{i=1}^{n}\left[\frac{-y_i(w^TX_i+b)}{-C-\gamma} - \frac{\gamma}{-C-\gamma}\right]^+ \\
s.t. \quad & \|w\| = 1.
\end{aligned}
\tag{39}
$$

We then make the change of variable $w_{new} = \frac{w}{C}, b_{new} = \frac{b}{C}$. Note then that

$$\begin{aligned} \frac{-y_i(w^T X_i + b)}{-C-\gamma} &= C\frac{-y_i(w_{new}^T X_i + b_{new})}{-C-\gamma} \\ &= \frac{-y_i(w_{new}^T X_i + b_{new})}{\|w_{new}\|(-C-\gamma)}. \end{aligned}$$

Plugging this and the new variables into (41), we get:

$$\begin{aligned} \min_{\gamma < -C, w_{new}, b_{new}} \quad & \sum_{i=1}^{n} \left[\frac{-y_i(w_{new}^T X_i + b_{new})}{\|w_{new}\|(-C-\gamma)} - \frac{\gamma}{-C-\gamma} \right]^+ \\ s.t. \quad & \|w_{new}\| = \frac{1}{C}. \end{aligned} \tag{40}$$

Finally, note that

$$\begin{aligned} \frac{1}{\|w_{new}\|(-C-\gamma)} + 1 &= \frac{1 + \|w_{new}\|(-C-\gamma)}{\|w_{new}\|(-C-\gamma)} \\ &= \frac{1 + (-1 - \frac{\gamma}{C})}{(-1 - \frac{\gamma}{C})} \\ &= \frac{-\gamma}{C} \left(\frac{C}{-C-\gamma} \right) \\ &= \frac{-\gamma}{-C-\gamma}. \end{aligned}$$

Plugging this into (40), we finally arrive at

$$\begin{aligned} \min_{\gamma < -C, w_{new}, b_{new}} \quad & \sum_{i=1}^{n} \left[\left(\frac{1}{-C-\gamma} \right) \left(\frac{-y_i(w_{new}^T X_i + b_{new}) + 1}{\|w_{new}\|} \right) + 1 \right]^+ \\ s.t. \quad & \frac{1}{\|w_{new}\|} = C. \end{aligned} \tag{41}$$

Appendix F. Bregman Divergence

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, let $\partial f(w)$ denote the set of subderivatives $\nabla f(w) \in \partial f(w)$ at a point $w \in \mathbb{R}^n$. For a differentiable, convex f the Bregman Divergence at $w_1, w_2 \in \mathbb{R}^n$ is defined as

$$d_f(w_1, w_2) = f(w_1) - f(w_2) - \langle \nabla f(w_2), w_1 - w_2 \rangle \, . \tag{42}$$

Since f is differentiable, $\nabla f(w)$ is the unique subderivative, and thus the divergence is well defined. If f is convex, but not differentiable, $\nabla f(w)$ may not be unique. In this case, however, following Appendix C of Bousquet and Elisseeff (2002), we can define a generalized Bregman Divergence. Letting the function $f^*(a) = \sup_w \langle w, a \rangle - f(w)$ denote the conjugate of f, we define the generalized Bregman Divergence at $w_1, a \in \mathbb{R}^n$ as,

$$\bar{d}_f(w_1, a) = f(w_1) + f^*(a) - \langle w_1, a \rangle \, . \tag{43}$$

As it relates to the normal Bregman Divergence, it is easy to check that for convex differentiable f,

$$d_f(w_1, w_2) = \bar{d}_f(w_1, \nabla f(w_2)) \ . \tag{44}$$

which follows from the property that

$$a \in \partial f(w) \iff f^*(a) = \langle a, w \rangle - f(w) \ .$$

As for its other properties, first note that the generalized divergence is non-negative. Second, notice that we have linearity, such that if $f = g+h$, for convex functions g, h, and we choose subderivatives of f, g, h satisfying $\nabla f(w_2) = \nabla g(w_2) + \nabla h(w_2)$, then

$$\bar{d}_f(w_1, \nabla f(w_2)) = \bar{d}_g(w_1, \nabla g(w_2)) + \bar{d}_h(w_1, \nabla h(w_2) \ . \tag{45}$$

To see this, we simply need to expand the right hand side in the following manner, where we use the fact that $a \in \partial f(w) \iff f^*(a) = \langle a, w \rangle - f(w)$ in the third and fifth equality.

$$\begin{aligned}
\bar{d}_f(w_1, \nabla f(w_2)) &= \bar{d}_{(g+h)}(w_1, \nabla (g+h)(w_2)) \\
&= (g+h)(w_1) + (g+h)^*(\nabla (g+h)(w_2)) - \langle w_1, \nabla (g+h)(w_2) \rangle \\
&= [(g+h)(w_1) - \langle w_1, \nabla (g+h)(w_2) \rangle] + \langle w_2, \nabla (g+h)(w_2) \rangle - (g+h)(w_2) \\
&= [(g+h)(w_1) - \langle w_1, \nabla (g+h)(w_2) \rangle] + (\langle w_2, \nabla g(w_2) \rangle - g(w_2)) \\
&\qquad\qquad + (\langle w_2, \nabla h(w_2) \rangle - h(w_2)) \\
&= [(g+h)(w_1) - \langle w_1, \nabla (g+h)(w_2) \rangle] + g^*(\nabla g(w_2)) + h^*(\nabla h(w_2)) \\
&= (g(w_1) + g^*(\nabla g(w_2)) - \langle w_1, \nabla g(w_2) \rangle) \\
&\qquad\qquad + (h(w_1) + h^*(\nabla h(w_2)) - \langle w_1, \nabla h(w_2) \rangle) \\
&= \bar{d}_g(w_1, \nabla g(w_2)) + \bar{d}_h(w_1, \nabla h(w_2) \ .
\end{aligned}$$

Finally, assume that w_f is a minimizer of f, meaning that $0 \in \partial f(w_f)$. Then, we have for any w_1,

$$\bar{d}_f(w_1, 0) = f(w_1) - f(w_f) \ . \tag{46}$$

References

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and trends® in machine learning*, 3(1):1–122, 2011.

Chih-Chung Chang and Chih-Jen Lin. Training v-support vector classifiers: theory and algorithms. *Neural computation*, 13(9):2119–2147, 2001.

Olivier Chapelle. Training a support vector machine in the primal. *Neural computation*, 19 (5):1155–1178, 2007.

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

Justin Davis and Stan Uryasev. Analysis of hurricane damage using buffered probability of exceedance. *Research report 2014-4, ise dept., university of florida*, 2014.

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.

Alexander Mafusalov and Stan Uryasev. Buffered probability of exceedance: Mathematical properties and optimization algorithms. *Research report 2014-1, ise dept., university of florida*, 2015.

Matthew Norton and Stan Uryasev. Maximization of auc and buffered auc in binary classification. *Research report 2014-2, ise dept., university of florida*, 2014.

Matthew Norton, Alexander Mafusalov, and Stan Uryasev. Cardinality of upper average and its application to network optimization. *Research report 2015-1, ise dept., university of florida*, 2015.

Fernando Pérez-Cruz, Jason Weston, DJL Herrmann, and Bernhard Schölkopf. Extension of the nu-svm range for classification. *Nato science series sub series III computer and system sciences*, 190:179–196, 2003.

Ralph Tyrell Rockafellar. Safeguarding strategies in risky optimization. *Presentation at the international workshop on engineering risk control and optimization, Gainesville, FL,*, 2009.

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

Ralph Tyrell Rockafellar and Johannes Royset. On buffered failure probability in design and optimization of structures. *Reliability engineering & system safety, Vol. 95, 499-510*, 2010.

Ralph Tyrell Rockafellar and Stan Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.

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

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

Akiko Takeda and Masashi Sugiyama. ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the 25th international conference on machine learning*, pages 1056–1063. ACM, 2008.

Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997.

Stan Uryasev. Buffered probability of exceedance and buffered service level: Definitions and properties. *Research report 2014-3, ise dept., university of florida*, 2014.

Ximing Wang, Neng Fan, and Panos M Pardalos. Stochastic subgradient descent method for large-scale robust chance-constrained support vector machines. *Optimization letters*, pages 1–12, 2016.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(Jul):1485–1510, 2009.