Robustness, Risk, and Regularization in Support Vector Machines

Huan Xu xuhuan@cim.mcgill.ca,
Department of Electrical and Computer Engineering, McGill University, Canada

Shie Mannor shie.mannor@mcgill.ca
Department of Electrical and Computer Engineering, McGill University, Canada

Constantine Caramanis cmcaram@ece.utexas.edu
Department of Electrical and Computer Engineering, The University of Texas at Austin, Texas, USA

Editor:

Abstract

We consider two new formulations for classification problems in the spirit of support vector machines based on robust optimization. Our formulations are designed to build in protection to noise and control overfitting, but without being overly conservative. Our first formulation allows the noise between different samples to be correlated. We show that the standard norm-regularized support vector machine classifier is a solution to a special case of our first formulation, thus providing an explicit link between regularization and robustness in pattern classification. Our second formulation is based on a softer version of robust optimization called comprehensive robustness. We show that this formulation is equivalent to regularization by any arbitrary convex regularizer, thus extending our first equivalence result. Moreover, we explain how the connection of comprehensive robustness to convex risk-measures can be used to design risk-measure constrained classifiers with robustness to the input distribution. Our formulations result in convex optimization problems that can be easily solved. Finally, we provide some empirical results that show the promise of comprehensive robust classifiers.

Keywords: Robustness, Risk, Regularization, Generalization, Support Vector Machine

1. Introduction

Support Vector Machines (SVMs for short), originated in Boser et al. (1992) and can be traced back to as early as Vapnik and Lerner (1963) and Vapnik and Chervonenkis (1974). They continue to be one of the most successful algorithms for classification. SVMs address the classification problem by finding the hyperplane (in the feature space) that achieves maximum sample margin when the training samples are separable. When the samples are not separable, a penalty term that approximates the total training-error is added to the minimizing objective (Bennett and Mangasrian, 1992; Cortes and Vapnik, 1995). It is well known that minimizing the training error itself can lead to poor classification performance for new unlabeled data; that is, such an approach may have poor generalization error because of, essentially, overfitting (Vapnik and Chervonenkis, 1991). A variety of modifications have been proposed to combat this problem, one of the most popular methods being that

of minimizing a combination of the training-error and a regularization term. The latter is typically chosen as a norm of the classifier. The resulting regularized classifier performs better on new data. This phenomenon is often interpreted from a statistical learning theory view: the regularization term restricts the complexity of the classifier, hence the deviation of the testing error and the training error is controlled (see Smola et al., 1998; Evgeniou et al., 2000; Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002; Bartlett et al., 2005, and references therein).

In this paper we follow a different approach, first proposed by Bhattacharyya et al. (2004b). We assume that the training data are generated by the true underlying distribution, but some non-iid (potentially adversarial) disturbance is then added to the samples we observe. We harness new developments in robust optimization (see Ghaoui and Lebret, 1997; Ben-Tal and Nemirovski, 1999; Bertsimas and Sim, 2004, and references therein), so-called comprehensive robust optimization (Ben-Tal et al., 2006b), and risk theory (Fóllmer and Schied, 2002; Ben-Tal et al., 2006a), to derive new robust SVM classifiers. The use of robust optimization in classification is not new (e.g., Shivaswamy et al., 2006; Bhattacharyya et al., 2004b; Lanckriet et al., 2002). Robust classification models studied in past work have considered only box-type uncertainty sets, which allow the possibility that the data have all been skewed in some non-neutral manner by a correlated disturbance. This has made it difficult to obtain non-conservative generalization bounds. Moreover, there has not been an explicit connection to the regularized classifier, although at a high-level it is known that regularization and robust optimization are related (e.g., Ghaoui and Lebret, 1997). The main contribution in this paper is the development of two new robust SVM classifiers that mitigate conservatism, provide an explicit connection to regularization (and as a byproduct PAC-style generalization error bounds), and provide the structure for efficiently computable classifiers satisfying risk measure constraints. In particular, our contributions include the following:

- Our first robust SVM formulation permits finer control of the adversarial disturbance, restricting it to satisfy aggregate constraints across data points, therefore reducing the possibility of highly correlated disturbance. This allows us to obtain bounds on the generalization error of the robust classifiers, as we show that as a special case of our robust formulation, we recover norm-based regularizers. In particular, we show that the norm-regularized SVM classifier is *equivalent* to a robust SVM classifier. We comment on this further below in this section.

- We next show that this new robust formulation is useful beyond complexity estimates and the precise connection to regularization: we use it to obtain considerably less conservative chance constraints, and we also use it to reprove consistency of SVM for classification.

- The second of our robust SVM formulations uses comprehensive robustness to construct "soft robust" classifiers whose performance is given different guarantees, based on the level of disturbance affecting the training data. This is in contrast to robust optimization, which provides the same guarantees uniformly inside the uncertainty set, and no guarantees outside. We show that this richer class of robustness is exactly equivalent to a much broader class of regularizers, including, e.g., KL divergence based

SVM regularizers, thus extending the scope of the previous equivalence. Moreover, we give favorable computational complexity results for these comprehensive robust classifiers.

- We next show the connection to risk theory, at the same time extending past work on chance constraints, and also opening the door for constructing classifiers with different risk-based guarantees. Although the connection seems natural, to the best of our knowledge this is the first attempt to view classification from a risk-hedging perspective.

In the final section, we illustrate the performance of our new classifiers through simulation. In particular we show that the comprehensive robust classifier, which can be viewed as a generalization of the standard SVM and the robust SVM, provides superior empirical results.

Robustness and Regularization

We comment here on our first result mentioned above, namely, the explicit equivalence of robustification and regularization. We briefly explain how this observation is different from previous work and why it is interesting. Indeed, certain equivalence relationships between robustification and regularization have been established for problems outside the machine learning field (Ghaoui and Lebret, 1997; Ben-Tal and Nemirovski, 1999), but their results do not directly apply to the classification problem. Indeed, research work on classifier regularization mainly discussed its effect on bounding the complexity of the function class (e.g., Smola et al., 1998; Evgeniou et al., 2000; Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002; Bartlett et al., 2005). Meanwhile, research work on robust classification has not attempted to relate robustness and regularization (e.g., Lanckriet et al., 2002; Bhattacharyya et al., 2004a,b; Shivaswamy et al., 2006; Trafalis and Gilbert, 2007; Globerson and Roweis, 2006), in part due to the robustness formulations used in those papers. (In fact, they all considered robustified versions of *regularized* classifications.)

The connection of robustification and regularization in the SVM context is important for the following reasons. First, it gives an alternative and potentially powerful explanation for the generalization ability of the regularization term. In the classical literature, the regularization term bounds the complexity of the class of classifiers. In our robust view of regularization, we regard the testing samples as a perturbed copy of the training samples. We then show that when the total perturbation is given or bounded, the regularization term bounds the gap between the classification errors of the SVM on these two sets of samples. In contrast to the PAC approach, this bound depends neither on how rich the class of candidate classifiers is, nor on an assumption that all samples are picked in an i.i.d. manner. In addition, this suggests novel approaches to designing good classification algorithms, in particular, designing the regularization term. In the PAC structural-risk minimization approach, regularization is chosen to minimize a bound on the generalization error based on the training error and a complexity term. This complexity term typically leads to overemphasizing the regularizer, and indeed this approach is known to often be too pessimistic (Kearns et al., 1997). The robust approach offers another avenue. Since both noise and robustness are physical processes, a close investigation of the application and noise characteristics at hand, can provide insights into how to properly robustify, and

therefore choose a regularization for the classifier. For example, it is widely known that normalizing the samples so that the variance among all features are roughly the same often leads to good generalization performance. From the robustness perspective, this simply says that the noise is skewed (ellipsoidal) rather than spherical, and hence an appropriate robustification must be designed to fit the skew of the physical noise process.

We also show that using the robust optimization viewpoint, we obtain some probabilistic results outside of the PAC setup. In Section 2.2 and Section 3.4 we bound the probability that a noisy training sample is correctly labeled. These bounds are different from the PAC bounds in that they consider the behavior of the *corrupted* samples. This is helpful when the training samples and the testing samples are drawn from different distributions, or some adversary manipulates the samples to prevent them from being correctly labeled. Finally, this connection of robustification and regularization also provides us with new proof techniques as well (see Section 2.3 for example).

Structure of the Paper: This paper is organized as follows. In Section 2 we investigate the correlated disturbance case, and show the equivalence between the robust classification and the regularization process. We also develop the connection to chance constraints. In Section 3 we investigate the comprehensive robust classification framework. We relate comprehensive robust classification with convex risk theory in Section 4. The kernelized version of comprehensive robust classification is given in Section 5. We provide numerical simulation results comparing robust classification and comprehensive robust classification in Section 6. Some concluding remarks are given in Section 7.

Notation: Capital letters are used to denote matrices, and boldface letters are used to denote column vectors. For a given norm $\|\cdot\|$, we use $\|\cdot\|^*$ to denote its dual norm. Similarly, for a function $f(\cdot)$ defined on a set $\mathcal{H}$, $f^*(\cdot)$ denotes its conjugate function, i.e., $f^*(\mathbf{y}) = \sup_{\mathbf{x}\in\mathcal{H}}\{\mathbf{y}^\top\mathbf{x} - f(\mathbf{x})\}$. For a vector $\mathbf{x}$ and a positive semi-definite matrix C of the same dimension, $\|\mathbf{x}\|_C$ denotes $\sqrt{\mathbf{x}^\top C\mathbf{x}}$. We use δ to denote disturbance affecting the samples. We use superscript r to denote the true value for an uncertain variable, so that $\boldsymbol{\delta}_i^r$ is the true (but unknown) noise of the i^{th} sample. The set of non-negative scalars is denoted by $\mathbb{R}^+$. The set of integers from 1 to n is denoted by $[1:n]$.

2. Robust Classification and Regularization

The main contributions of this section are: (i) we formulate and solve a new robust classification problem which, unlike previous formulations, is able to limit or constrain the adversary to using a correlated disturbance; (ii) using this model, we show that the standard regularized classifier is a special case of our robust classification, thus explicitly relating robustness and regularization. This provides an alternative explanation for the success of regularization, and also suggests new physically-motivated ways to construct regularizers; (iii) we formulate a chance-constrained classifier which can be approximated by the robust formulation for correlated disturbance, and as a result is far less conservative than what previous models could provide; (iv) finally, we show that the robustness perspective can be useful in its own right, by using it to prove a consistency result for regularized SVM classification.

2.1 Robust Classification for Correlated Disturbance

We consider the standard 2-class classification setup, where we are given a finite number of training samples $\{\mathbf{x}_i, y_i\}_{i=1}^m \subseteq \mathbb{R}^n \times \{-1, +1\}$, and must find a linear classifier, specified by the function $h^{\mathbf{w},b}(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$. For the standard regularized classifier, the parameters $(\mathbf{w}, b)$ are obtained by solving the following convex optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b} : \quad & r(\mathbf{w}, b) + \sum_{i=1}^m \xi_i \\ \text{s.t.} : \quad & \xi_i \geq \left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\right] \\ & \xi_i \geq 0, \end{aligned}$$

where $r(\mathbf{w}, b)$ is a regularization term. The standard robust optimization techniques robustify at a constraint-wise level, allowing the disturbances $\vec{\boldsymbol{\delta}} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m)$ to lie in some uncertainty set $\mathcal{N}$:

$$\begin{aligned} \min_{\mathbf{w},b} : \quad & r(\mathbf{w}, b) + \sum_{i=1}^m \xi_i \\ \text{s.t.} : \quad & \xi_i \geq \left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b)\right], \quad \vec{\boldsymbol{\delta}} \in \mathcal{N}, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

It is well-known (e.g., Ben-Tal et al., 2003) that due to the constraint-wise uncertainty formulation, the uncertainty set is effectively rectangular; that is, if $\mathcal{N}_i$ denotes the projection of $\mathcal{N}$ onto the $\boldsymbol{\delta}_i$ component, then replacing $\mathcal{N}$ by the potentially larger product set $\mathcal{N}_{\text{box}} = \mathcal{N}_1 \times \cdots \times \mathcal{N}_m$ yields an equivalent formulation. Effectively, this allows simultaneous worst-case disturbances across many constraints, and this is exactly what leads to overly conservative formulations. The goal of this paper is to obtain a robust formulation where the disturbances $\{\boldsymbol{\delta}_i\}$ may be meaningfully taken to be correlated, so that the problem is no longer equivalent to the box case. In order to side-step this problem, we robustify an equivalent SVM formulation:

$$\min_{\mathbf{w},b} \left\{ r(\mathbf{w}, b) + \sum_{i=1}^m \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b), 0\right] \right\},$$

and we thus obtain:

$$\min_{\mathbf{w},b} \max_{\vec{\boldsymbol{\delta}} \in \mathcal{N}} \left\{ r(\mathbf{w}, b) + \sum_{i=1}^m \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b), 0\right] \right\}. \tag{2}$$

Note that the problem (1) above is equivalent to:

$$\min_{\mathbf{w},b} \max_{\vec{\boldsymbol{\delta}} \in \mathcal{N}_{\text{box}}} \left\{ r(\mathbf{w}, b) + \sum_{i=1}^m \max\left[1 - y_i(\langle \mathbf{w}\, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b), 0\right] \right\}. \tag{3}$$

We define explicitly the correlated disturbance (or uncertainty) set to be investigated.

Definition 1 *1. A set $\mathcal{N}_0 \subseteq \mathbb{R}^n$ is called an* Atomic Uncertainty Set *if*

$$\begin{aligned} &(I) \quad \mathbf{0} \in \mathcal{N}_0; \\ &(II) \quad \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} \left[\mathbf{w}^\top \boldsymbol{\delta}\right] = \sup_{\boldsymbol{\delta}' \in \mathcal{N}_0} \left[-\mathbf{w}^\top \boldsymbol{\delta}'\right] < \infty, \ \forall \mathbf{w} \in \mathbb{R}^n. \end{aligned}$$

2. Let $\mathcal{N}_0$ be an atomic uncertainty set. A set $\mathcal{N} \subseteq \mathbb{R}^{n \times m}$ is called a Concave Correlated Uncertainty Set *of $\mathcal{N}_0$, if*

$$\begin{aligned} &(I) \quad \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \boldsymbol{\delta}_t \in \mathcal{N}_0; \ \boldsymbol{\delta}_{i \neq t} = \mathbf{0}\} \subseteq \mathcal{N}, \ \forall\, t; \\ &(II) \quad \mathcal{N} \subseteq \{(\alpha_1 \boldsymbol{\delta}_1, \cdots, \alpha_m \boldsymbol{\delta}_m) | \sum_{i=1}^m \alpha_i = 1; \ \alpha_i \geq 0, \ \boldsymbol{\delta}_i \in \mathcal{N}_0, \ \forall\, i\}. \end{aligned}$$

3. Conversely, given a set $\mathcal{N}$ we define the corresponding atomic uncertainty set: For $\mathcal{N} \subseteq \mathbb{R}^{n \times m}$ an uncertainty set, its atomic uncertainty set $\mathcal{N}_0 \subseteq \mathbb{R}^n$ is the smallest set such that

$$\mathcal{N} \subseteq \{(\alpha_1 \boldsymbol{\delta}_1, \cdots, \alpha_m \boldsymbol{\delta}_m) | \sum_{i=1}^m \alpha_i = 1; \ \alpha_i \geq 0, \ \boldsymbol{\delta}_i \in \mathcal{N}_0, \ \forall\, i\},$$

i.e., it is contained in the largest Concave Correlated Uncertainty set of $\mathcal{N}_0$. The significance of this final definition becomes apparent in the theorem below.

The concave correlated uncertainty definition models the case where the disturbances on each sample are treated identically, but their aggregate behavior across multiple samples is controlled. Some interesting examples include

$$\begin{aligned} &(1) \quad \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \sum_{i=1}^m \|\boldsymbol{\delta}_i\| \leq c\}; \\ &(2) \quad \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \exists t \in [1:m]; \ \|\boldsymbol{\delta}_t\| \leq c, \ \boldsymbol{\delta}_{i \neq t} = \mathbf{0}\}; \\ &(3) \quad \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \sum_{i=1}^m \sqrt{c\|\boldsymbol{\delta}_i\|} \leq c\}. \end{aligned}$$

Theorem 2 *Assume $\{\mathbf{x}_i, y_i\}_{i=1}^m$ are non-separable, $r(\cdot) : \mathbb{R}^{n+1} \to \mathbb{R}$ is an arbitrary function, $\mathcal{N}$ is a concave correlated uncertainty set with corresponding atomic uncertainty set $\mathcal{N}_0$. Then the following min-max problem*

$$\inf_{\mathbf{w}, b} \sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}} \left\{ r(\mathbf{w}, b) + \sum_{i=1}^m \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b), \, 0\right] \right\} \tag{4}$$

is equivalent to

$$\begin{aligned} \min: \quad & r(\mathbf{w}, b) + \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} (\mathbf{w}^\top \boldsymbol{\delta}) + \sum_{i=1}^m \xi_i, \\ \text{s.t.}: \quad & \xi_i \geq 1 - [y_i(\langle \mathbf{w}, \, \mathbf{x}_i \rangle + b)], \quad i = 1, \cdots, m; \\ & \xi_i \geq 0, \quad i = 1, \cdots, m. \end{aligned} \tag{5}$$

Furthermore, the minimization of Problem (5) is attainable when $r(\cdot, \cdot)$ is lower semi-continuous.

Proof We first make the following definitions:

$$\begin{aligned}
\mathcal{N}^- &\triangleq \bigcup_{t=1}^{m} \mathcal{N}_t^-; \quad \text{where:} \quad \mathcal{N}_t^- \triangleq \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \boldsymbol{\delta}_t \in \mathcal{N}_0;\ \boldsymbol{\delta}_{i \neq t} = \mathbf{0}\}. \\
\mathcal{N}^+ &\triangleq \{(\alpha_1 \boldsymbol{\delta}_1, \cdots, \alpha_m \boldsymbol{\delta}_m) | \sum_{i=1}^{m} \alpha_i = 1;\ \alpha_i \geq 0,\ \boldsymbol{\delta}_i \in \mathcal{N}_0,\ i = 1, \cdots, m\}; \\
v(\mathbf{w}, b) &\triangleq \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} (\mathbf{w}^\top \boldsymbol{\delta}) + \sum_{i=1}^{m} \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b),\, 0\right].
\end{aligned}$$

Recall that $\mathcal{N}^- \subseteq \mathcal{N} \subseteq N^+$ by definition. Hence, fixing any $(\hat{\mathbf{w}}, \hat{b}) \in \mathbb{R}^{n+1}$, the following inequalities hold:

$$\begin{aligned}
&\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}^-} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\, 0\right] \\
\leq &\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\, 0\right] \\
\leq &\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}^+} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\, 0\right].
\end{aligned}$$

To prove the theorem, we first show that $v(\hat{\mathbf{w}}, \hat{b})$ is no larger than the leftmost expression and then show $v(\hat{\mathbf{w}}, \hat{b})$ is no smaller than the rightmost expression.
Step 1: First we prove

$$v(\hat{\mathbf{w}}, \hat{b}) \leq \sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}^-} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\, 0\right]. \tag{6}$$

Since the samples $\{\mathbf{x}_i,\, y_i\}_{i=1}^m$ are not separable, there exists $t \in [1:m]$ such that

$$y_t(\langle \hat{\mathbf{w}}, \mathbf{x}_t \rangle + \hat{b}) < 0. \tag{7}$$

Hence,

$$\begin{aligned}
&\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}_t^-} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\, 0\right] \\
&= \sum_{i \neq t} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}),\, 0\right] + \sup_{\boldsymbol{\delta}_t \in \mathcal{N}_0} \max\left[1 - y_t(\langle \hat{\mathbf{w}}, \mathbf{x}_t - \boldsymbol{\delta}_t \rangle + \hat{b}),\, 0\right] \\
&= \sum_{i \neq t} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}),\, 0\right] + \max\left[1 - y_t(\langle \hat{\mathbf{w}}, \mathbf{x}_t \rangle + \hat{b}) + \sup_{\boldsymbol{\delta}_t \in \mathcal{N}_0} (y_t \hat{\mathbf{w}}^\top \boldsymbol{\delta}_t),\, 0\right] \\
&= \sum_{i \neq t} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}),\, 0\right] + \max\left[1 - y_t(\langle \hat{\mathbf{w}}, \mathbf{x}_t \rangle + \hat{b}),\, 0\right] + \sup_{\boldsymbol{\delta}_t \in \mathcal{N}_0} (y_t \hat{\mathbf{w}}^\top \boldsymbol{\delta}_t) \\
&= \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} (\hat{\mathbf{w}}^\top \boldsymbol{\delta}) + \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}), 0\right] \\
&= v(\hat{\mathbf{w}}, \hat{b}).
\end{aligned}$$

The third equality holds because of Inequality (7) and $\sup_{\boldsymbol{\delta}_t \in \mathcal{N}_0}(y_t \hat{\mathbf{w}}^\top \boldsymbol{\delta}_t)$ being non-negative (recall $\mathbf{0} \in \mathcal{N}_0$). Since $\mathcal{N}_t^- \subseteq \mathcal{N}^-$, Inequality (6) follows.

Step 2: Next we prove

$$\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}^+} \sum_{i=1}^m \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\, 0\right] \leq v(\hat{\mathbf{w}}, \hat{b}). \tag{8}$$

Notice that by the definition of $\mathcal{N}^+$ we have

$$\begin{aligned}
&\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}^+} \sum_{i=1}^m \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\, 0\right] \\
=& \sup_{\sum_{i=1}^m \alpha_i = 1;\, \alpha_i \geq 0;\, \hat{\boldsymbol{\delta}}_i \in \mathcal{N}_0} \sum_{i=1}^m \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \alpha_i \hat{\boldsymbol{\delta}}_i \rangle + \hat{b}),\, 0\right] \qquad (9) \\
=& \sup_{\sum_{i=1}^m \alpha_i = 1;\, \alpha_i \geq 0;} \sum_{i=1}^m \max\Big[\sup_{\hat{\boldsymbol{\delta}}_i \in \mathcal{N}_0} \big(1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \alpha_i \hat{\boldsymbol{\delta}}_i \rangle + \hat{b})\big),\, 0\Big].
\end{aligned}$$

Now, for any $i \in [1:m]$, the following holds,

$$\begin{aligned}
&\max\Big[\sup_{\hat{\boldsymbol{\delta}}_i \in \mathcal{N}_0} \big(1 - y_i(\langle \hat{\mathbf{w}},\, \mathbf{x}_i - \alpha_i \hat{\boldsymbol{\delta}}_i \rangle + \hat{b})\big),\, 0\Big] \\
=& \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) + \alpha_i \sup_{\hat{\boldsymbol{\delta}}_i \in \mathcal{N}_0} (\hat{\mathbf{w}}^\top \hat{\boldsymbol{\delta}}_i),\, 0\right] \\
\leq& \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}),\, 0\right] + \alpha_i \sup_{\hat{\boldsymbol{\delta}}_i \in \mathcal{N}_0} (\hat{\mathbf{w}}^\top \hat{\boldsymbol{\delta}}_i).
\end{aligned}$$

Therefore, Equation (9) is upper bounded by

$$\begin{aligned}
&\sum_{i=1}^m \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}),\, 0\right] + \sup_{\sum_{i=1}^m \alpha_i = 1;\, \alpha_i \geq 0;} \sum_{i=1}^m \alpha_i \sup_{\hat{\boldsymbol{\delta}}_i \in \mathcal{N}_0} (\hat{\mathbf{w}}^\top \hat{\boldsymbol{\delta}}_i) \\
=& \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} (\hat{\mathbf{w}}^\top \boldsymbol{\delta}) + \sum_{i=1}^m \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}), 0\right] \\
=& \, v(\hat{\mathbf{w}}, \hat{b}),
\end{aligned}$$

hence Inequality (8) holds.

Combining the two steps and adding $r(\mathbf{w}, b)$ on both side leads to: $\forall (\mathbf{w}, b) \in \mathbb{R}^{n+1}$,

$$\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \hat{\mathcal{N}}} \sum_{i=1}^m \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b),\, 0\right] + r(\mathbf{w}, b) = v(\mathbf{w}, b) + r(\mathbf{w}, \mathbf{b}).$$

Taking the infimum on both sides establishes the equivalence of Problem (4) and Problem (5). Observe that $\sup_{\boldsymbol{\delta} \in \mathcal{N}_0} \mathbf{w}^\top \boldsymbol{\delta}$ is a supremum over a class of affine functions, and hence is lower semi-continuous. Therefore $v(\cdot, \cdot)$ is also lower semi-continuous. Thus the minimum can be achieved for Problem (5), and Problem (4) by equivalence, when $r(\cdot)$ is lower semi-continuous. ■

This theorem reveals the main difference in conservatism between the constraint-wise uncertainty in (1) and our formulation in (2). Consider both formulations with the same uncertainty set, $\mathcal{N} = \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \sum_{i=1}^m \|\boldsymbol{\delta}_i\| \leq c\}$. The corresponding atomic set of $\mathcal{N}$ is $\mathcal{N}_0 = \{\|\delta\| \leq c\}$, but the atomic set for $\mathcal{N}_{\text{box}}$ is $m\mathcal{N}_0$ since $\mathcal{N}_{\text{box}} = \mathcal{N}_0 \times \cdots \times \mathcal{N}_0$ includes disturbances with a magnitude as large as mc. Therefore the latter (recall (3)) is equivalent to a regularization coefficient of the form $m\lambda$, that is linked to the number of training samples, and will therefore be overly conservative.

An immediate corollary is that a special case of our robust formulation is exactly equivalent to the norm-regularized SVM setup:

Corollary 3 *Let* $\mathcal{T}_k \triangleq \Big\{(\boldsymbol{\delta}_1, \cdots \boldsymbol{\delta}_m) | \sum_{i=1}^m \|\boldsymbol{\delta}_i\| \leq c;\ \#\{i|\boldsymbol{\delta}_i = \mathbf{0}\} \geq m - k\Big\}$, [1] *for* $k \in [1 : m]$ *and* $c > 0$. *If the training sample* $\{\mathbf{x}_i, y_i\}_{i=1}^m$ *are non-separable, then the following two optimization problems on* $(\mathbf{w}, b)$ *are equivalent*

$$\min: \qquad \max_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) \in \mathcal{T}_k} \sum_{i=1}^m \max\big[1 - y_i\big(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b\big), 0\big], \tag{10}$$

$$\min: \qquad c\|\mathbf{w}\|^* + \sum_{i=1}^m \max\big[1 - y_i\big(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b\big), 0\big]. \tag{11}$$

Proof Let $\mathcal{N}_0$ be the norm-ball and $r(\mathbf{w}, b) \equiv 0$. Then $\sup_{\|\boldsymbol{\delta}\| \leq c}(\mathbf{w}^\top \boldsymbol{\delta}) = c\|\mathbf{w}\|^*$. The corollary follows from Theorem 2. ■

This corollary explains the widely known fact that the regularized classifier tends to be more robust. Specifically, it explains the observation that when the disturbance is noise-like and neutral rather than adversarial, a norm-regularized classifier (without any robustness requirement) has a performance often superior to the *constraint-wise* robust classifier (see Trafalis and Gilbert, 2007, and Section 6 of this paper). On the other hand, this observation also suggests that the appropriate way to regularize should come from a disturbance-robustness perspective. The above equivalence implies that standard regularization essentially assumes that the disturbance is spherical; if this is not true, robustness may yield a better regularization-like algorithm. To find a more effective regularization term, a closer investigation of the data variation is desirable, e.g., by examining the variation of the data and solving the corresponding robust classification problem. For example, one way to regularize is by splitting the given training samples into two subsets with equal number of elements, and treating one as a disturbed copy of the other. By analyzing the direction of the disturbance and the magnitude of the total variation, one can choose the proper norm to use, and a suitable tradeoff parameter.

2.2 Probabilistic Interpretation

Although Problem (4) is formulated without any probabilistic assumptions, it can be used to approximate an upper bound for a chance-constrained classifier. Suppose the disturbance

1. $\mathcal{T}_k$ is the uncertainty set capturing the setting where no more than k samples can be disturbed, while the magnitude of the total disturbance is bounded.

$(\boldsymbol{\delta}_1^r, \cdots \boldsymbol{\delta}_m^r)$ follows a joint probability measure μ. Then the chance-constrained classifier is given by the following minimization problem on $(\mathbf{w}, b, l)$ given a confidence level $\eta \in [0,\, 1]$,

$$\begin{aligned} \min: \quad & l \\ \text{s.t.}: \quad & \mu\Big\{ \textstyle\sum_{i=1}^m \max\big[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i^r\rangle + b), 0\big] \le l \Big\} \ge 1 - \eta. \end{aligned} \tag{12}$$

The formulations in Shivaswamy et al. (2006); Lanckriet et al. (2002); Bhattacharyya et al. (2004a) assume uncorrelated noise and require all constraints to be satisfied with high probability *simultaneously*. They find a vector $[\xi_1, \cdots, \xi_m]^\top$ such that each ξ_i bounds the hinge-loss for sample $\mathbf{x}_i^r$ with high probability. In contrast, our formulation above bounds the average (or equivalently the sum of) empirical error. When controlling this average quantity is of more interest, the uncorrelated-noise formulation will be overly conservative.

Problem (12) is generally non-tractable. However, we can approximate it as follows. Let

$$c^* \triangleq \inf\{\alpha | \mu(\sum_i \|\boldsymbol{\delta}_i\| \le \alpha) \ge 1 - \eta\}.^2$$

Then for any $(\mathbf{w}, b)$, with probability no less than $1 - \eta$, the following holds,

$$\begin{aligned} & \sum_{i=1}^m \max\big[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i^r\rangle + b), 0\big] \\ \le & \max_{\sum_i \|\boldsymbol{\delta}_i\| \le c^*} \sum_{i=1}^m \big[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b), 0\big]. \end{aligned}$$

Thus (12) is upper bounded by (11) with $c = c^*$. This gives an additional probabilistic robustness property of the standard regularized classifier. Notice that following a similar approach but with the constraint-wise robust setup, i.e., the box uncertainty set, would lead to considerably more pessimistic approximations of the chance constraint.

2.3 Consistency of Regularization

In this subsection, we work out a simple example to illustrate how the robustness perspective might help in a statistical learning setup, by establishing the consistency of the linear classifier.

The following theorem is a well-known result in statistical machine learning (Steinwart, 2005). Here we reprove it using our robust classifier setup, by bounding the total variation between the set of test samples and the set of training samples.

Theorem 4 *Let P be the underlying generating probability with bounded support $\mathcal{X} \times \{-1, +1\}$, where $\mathcal{X} \subseteq \mathbb{R}^n$. Then for $c > 0$ there exists $\{\gamma_s\} \to 0$ independent of $(\mathbf{w},\, b)$ such that*

$$\mathbb{E}_{\mathbb{P}}(\mathbf{1}_{y \ne sgn(\langle \mathbf{w}, \mathbf{x}\rangle + b)}) \le \gamma_N + c\|\mathbf{w}\|_2 + \frac{1}{N}\sum_{i=1}^N \max\big[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b), 0\big],$$

holds almost surely as $N \to +\infty$.

2. Given μ, c^* is easily simulated, and for specific probability measures (e.g., independent Gaussian disturbance), it can be computed analytically.

Proof For $c > 0$, a testing sample $(\mathbf{x}', y')$ and a training sample $(\mathbf{x}, y)$ are called a *sample pair* if $y = y'$ and $\|\mathbf{x} - \mathbf{x}'\|_2 \leq c$. We say a set of training samples and a set of testing samples form l pairings if there exist l sample pairs with no data reused. Given n training samples and n testing samples, we use M_n to denote the largest number of pairings. To prove this theorem, we need to establish the following lemma.

Lemma 5 *Given $c > 0$, $M_n/n \to 1$ almost surely as $n \to +\infty$.*

Proof We make a partition of $\mathcal{X} \times \{-1, +1\} = \bigcup_{i=1}^{T} \mathcal{X}_i$ such that $\mathcal{X}_i$ either has the form $[\alpha_1, \alpha_1 + c/\sqrt{n}) \times [\alpha_2, \alpha_2 + c/\sqrt{n}) \cdots \times [\alpha_n, \alpha_n + c/\sqrt{n}) \times \{+1\}$ or $[\alpha_1, \alpha_1 + c/\sqrt{n}) \times [\alpha_2, \alpha_2 + c/\sqrt{n}) \cdots \times [\alpha_n, \alpha_n + c/\sqrt{n}) \times \{-1\}$. That is, each partition is the cartesian product of a rectangular cell in $\mathcal{X}$ and a singleton in $\{-1, +1\}$. Notice that if a training sample and a testing sample fall into $\mathcal{X}_i$, they can form a pairing.

Let $\mathbb{P}_n^{tr}$ and $\mathbb{P}_n^{te}$ be the empirical distribution of training samples and testing samples, respectively. Now we calculate the number of unpaired samples $n - M_n$. This can be upper bounded by

$$\sum_{i=1}^{T} |\#(\text{training samples in } \mathcal{X}_i) - \#(\text{testing samples in } \mathcal{X}_i)| = n \sum_{i=1}^{T} |\int I_{\mathcal{X}_i} d\mathbb{P}_n^{tr} - \int I_{\mathcal{X}_i} d\mathbb{P}_n^{te}|.$$

Furthermore, letting $\mathcal{F}$ be the set of indicator functions $I_{\mathcal{X}_i}$, then $\mathcal{F}$ is a P-Donsker class, and hence a Glivenko-Cantelli class almost surely. We thus have

$$\sup_{f \in \mathcal{F}} |\int f d\mathbb{P}_n^{tr} - \int f d\mathbb{P}_n^{te}| \to 0,$$

almost surely when $n \to +\infty$. This leads to

$$\sum_{i=1}^{T} |\int I_{\mathcal{X}_i} d\mathbb{P}_n^{tr} - \int I_{\mathcal{X}_i} d\mathbb{P}_n^{te}| \to 0.$$

Therefore $(n - M_n)/n \to 0$ almost surely. ■

Now we proceed to prove the theorem. Given n training samples and n testing samples with M_n sample pairs, we notice that for these paired samples, the total testing error is upper bounded by

$$\begin{aligned} &\max_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_n) \in \mathcal{T}_N} \sum_{i=1}^{n} \max\left[1 - y_i\big(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b\big), 0\right] \\ =& cn\|\mathbf{w}\|_2 + \sum_{i=1}^{n} \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b), 0\right]. \end{aligned}$$

Hence the classification error of the total n testing samples can be upper bounded by

$$(n - M_n) + cn\|\mathbf{w}\|_2 + \sum_{i=1}^{n} \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b), 0\right].$$

Therefore, the average testing error is upper bounded by

$$1 - M_n/n + c\|\mathbf{w}\|_2 + \frac{1}{n}\sum_{i=1}^{n} \max\left[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b),\, 0\right].$$

Notice that $M_n/n \to 1$ almost surely. ■

3. Comprehensive Robust Classification

Robust optimization provides a solution with but one guarantee: feasibility and worst-case performance control for any realization of the uncertainty within the bounded uncertainty set. If the uncertainty realization turns out favorable (e.g., close to mean behavior), no improved performance is guaranteed, while if the realization occurs outside the assumed uncertainty set, all bets are off: the error is not controlled. This characteristic makes it difficult to address noise with fat tails: if we take a small uncertainty set, we have no protection guarantees for potentially high probability events; on the other hand, if we seek to protect ourselves over large uncertainty sets, the robust setting may yield overly pessimistic solutions. In this section we address exactly this problem, by designing a new classifier with performance guarantees indexed to the level of noise. We use the softer notion of "comprehensive robustness," recently explored in the robust optimization literature (Ben-Tal et al., 2006b).

This allows us to construct classifiers with improved empirical performance. In addition, we show that this new notion of robustness yields a broader range of regularization schemes than robust optimization, including squared-norm, and Kullback-Leibler regularization. Moreover, extending the chance constraint results of the previous section, we are able to provide probability bounds for *all* magnitudes of constraint violations.

The key idea to comprehensive robustness is to discount lower-probability noise realizations, by reducing the loss incurred. If we denote the hinge loss of a sample under a certain noise realization as $\xi_i(\boldsymbol{\delta}_i) \triangleq \max\left[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b), 0\right]$, the robust classifier (2) can be rewritten as:

$$\min_{\mathbf{w},b} \max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\in\mathcal{N}} \left\{r(\mathbf{w},b) + \sum_{i=1}^{m} \xi_i(\boldsymbol{\delta}_i)\right\}.$$

Instead, we formulate the comprehensive robust classifier by introducing a discounted loss function depending not only on the nominal hinge loss, but also on the noise realization itself. Let $h_i(\cdot,\cdot) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ satisfy $0 \le h_i(\alpha,\beta) \le h_i(\alpha,\mathbf{0}) = \alpha$. We use h to denote our discounted loss function: it discounts the loss depending on the realized data, yet is always nonnegative, and provides no discount for samples with zero disturbance. Thus, the comprehensive robust classifier is given by:

$$\min_{\mathbf{w},b} \sup_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\in\mathcal{N}} \left\{r(\mathbf{w},b) + \sum_{i=1}^{m} h_i\big(\xi_i(\boldsymbol{\delta}_i), \boldsymbol{\delta}_i\big)\right\}. \tag{13}$$

We primarily investigate additive discounts of the form $h_i(\alpha,\boldsymbol{\beta}) \triangleq \max(0, \alpha - f_i(\boldsymbol{\beta}))$, taking a brief detour in Section 3.5 to consider multiplicative discounts. Additive structure

provides a rich class of discount functions, while remaining tractable. Moreover, additive structure provides the link to risk theory and convex risk measures, which we consider in Section 4.

We formulate the comprehensive robust classification with additive discount function in Section 3.1 and establish an equivalence relationship between comprehensive robust classification and a broad class of regularization schemes in Section 3.2. In particular, we show that the standard norm-regularized SVM has a comprehensive robust representation, and so do many regularized SVMs with non-norm regularizers.

In Section 3.3 we investigate the tractability of comprehensive robust classification. In Section 3.4 we discuss a special class of discounts, namely norm discounts, and derive probability bounds for such discounts. Finally, in Section 3.5 we briefly investigate the tractability of multiplicative discount functions with the form $h_i(\alpha, \boldsymbol{\beta}) \triangleq c(\boldsymbol{\beta}) \max(0, \alpha)$.

3.1 Problem Formulation

We consider box uncertainty sets throughout, to facilitate some of the analysis and allow focus on the effect of the discount function.[3] Substituting $h_i(\alpha, \boldsymbol{\beta}) \triangleq \max(0, \alpha - f_i(\boldsymbol{\beta}))$ into Equation (13) and extending $f_i(\cdot)$ to take the value $+\infty$ for $\delta_i \notin \mathcal{N}_i$, we have the formulation of the comprehensive robust classifier:

Comprehensive Robust Classifier:

$$\begin{aligned} \min: \quad & r(\mathbf{w}, b) + \sum_{i=1}^{m} \xi_i, \\ \text{s.t.}: \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \mathbb{R}^n,\ i = 1, \cdots, m \\ & \xi_i \geq 0; \qquad i = 1, \cdots, m. \end{aligned}$$

This $f_i(\cdot)$ (extended real) function controls the disturbance discount, and therefore must satisfy

$$\inf_{\boldsymbol{\beta} \in \mathbb{R}^n} f_i(\boldsymbol{\beta}) = f_i(\mathbf{0}) = 0. \tag{14}$$

Notice that if we set $f_i(\cdot)$ to be the indicator function of a set, we recover the standard robust classifier. Thus the comprehensive robust classifier is a natural generalization of the robust classifier with more flexibility on setting $f_i(\cdot)$.

The $f_i(\cdot)$ function also has a physical interpretation as controlling the margin of the resulting classifier under *all* noise. That is, when $\xi_i = 0$, the resulting classifier guarantees a margin $1/\|\mathbf{w}\|$ for the observed sample $\mathbf{x}_i$ (the same as the standard classifier), together with a guaranteed margin $(1 - f_i(\boldsymbol{\delta}_i))/\|\mathbf{w}\|$ when the sample is perturbed by $\boldsymbol{\delta}_i$.

3.2 Comprehensive Robustness and Regularization

In this section we show that any convex regularization term in the constraint is equivalent to a comprehensive robust formulation, and vice versa. Moreover, the standard regularized SVM is equivalent to a (non-regularized) comprehensive robust classifier where $f_i(\boldsymbol{\delta}_i) = \alpha\|\boldsymbol{\delta}_i\|$.

3. Nevertheless, we expect that combining the analysis of Section 2 will yield interesting results.

Given a function $f(\cdot)$, let f^* denote its Legendre-Fenchel transform or conjugate function, given by $f^*(s) = \sup_x\{\langle s, y\rangle - f(x)\}$ (Rockafellar, 1970). Then we have the following, that shows that if f is a disturbance discount that satisfies (14), then so does its conjugate, and vice versa. We use this below to establish the equivalence between convex regularization and comprehensive robustness.

Lemma 6 *(i) If $f(\cdot)$ satisfies (14), then so does $f^*(\cdot)$.*

(ii) If $g(\cdot)$ is closed and convex, and $g^*(\cdot)$ satisfies (14), then so does $g(\cdot)$.

Proof

(i) By definition we have $f^*(\mathbf{y}) \geq \mathbf{y}^\top \mathbf{0} - f(\mathbf{0})$, $\forall \mathbf{y} \in \mathbb{R}^n$. Hence $\inf_{\mathbf{y}\in\mathbf{R}^n} f^*(\mathbf{y}) \geq 0$, since $f(\mathbf{0}) = 0$. Furthermore, $f^*(\mathbf{0}) = \sup_{\mathbf{x}\in\mathbb{R}^n}(\mathbf{0}^\top\mathbf{x} - f(\mathbf{x})) = -\inf_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) = 0$ completes the proof of the first part.

(ii) For $g(\cdot)$ closed and convex, $g(\cdot) = (g(\cdot)^*)^*$ (Rockafellar, 1970; Boyd and Vandenberghe, 2004). The second part follows from the first part by setting $f(\cdot) = g^*(\cdot)$. ■

Theorem 7 *The Comprehensive Robust Classifier (14) is equivalent to the following convex program:*

$$\begin{aligned} \min :& r(\mathbf{w}, b) + \sum_{i=1}^m \xi_i, \\ \text{s.t. }:& y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - f_i^*(y_i\mathbf{w}) \geq 1 - \xi_i, \;\; i = 1, \cdots, m, \\ & \xi_i \geq 0, \;\; i = 1, \cdots, m. \end{aligned} \tag{15}$$

Proof Simple algebra yields

$$\begin{aligned} & y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \;\; \forall \boldsymbol{\delta}_i \in \mathbb{R}^n \\ \iff \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - y_i\mathbf{w}^\top\boldsymbol{\delta}_i + f_i(\boldsymbol{\delta}_i) \geq 1 - \xi_i, \;\; \forall \boldsymbol{\delta}_i \in \mathbb{R}^n \\ \iff \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - \sup_{\boldsymbol{\delta}_i\in\mathbb{R}^n} \left[y_i\mathbf{w}^\top\boldsymbol{\delta}_i - f_i(\boldsymbol{\delta}_i)\right] \geq 1 - \xi_i \\ \iff \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - f_i^*(y_i\mathbf{w}) \geq 1 - \xi_i. \end{aligned}$$

Finally, note that the problem convexity follows immediately from the (generic) convexity of the conjugate function. ■

From Lemma 6(i),

$$\inf_{\mathbf{w}\in\mathbb{R}^n} f_i^*(y_i\mathbf{w}) = f_i^*(\mathbf{0}) = 0,$$

and therefore $f_i^*(\cdot)$ "penalizes" $y_i\mathbf{w}$ and is thus a regularization term. On the other hand, a classifier that has a convex regularization term $g(\cdot)$ in each constraint is equivalent to a comprehensive robust classifier with disturbance discount $f(\cdot) = g^*(\cdot)$ (Lemma 6(ii)). Therefore, the comprehensive robust classifier is equivalent to the constraint-wise regularized classifier with general convex regularization. This equivalence gives an alternative explanation

for the generalization ability of regularization: intuitively, the set of testing data can be regarded as a "disturbed" copy of the set of training samples where the penalty on large (or low-probability) disturbance is discounted. Empirical results show that a classifier that handles noise well has a good performance for testing samples.

As an example of this equivalence, set $f_i(\boldsymbol{\delta}_i) = \alpha\|\boldsymbol{\delta}_i\|$ for $\alpha > 0$ and $r(\mathbf{w}, b) \equiv 0$. Here,

$$f_i^*(y_i\mathbf{w}) = \begin{cases} 0 & \|\mathbf{w}\|^* \leq \alpha, \\ +\infty & \text{otherwise;} \end{cases}$$

which is the indicator function of the dual-norm ball with radius α. Thus (15) is equivalent to

$$\begin{array}{ll} \min: & \sum_{i=1}^m \xi_i, \\ \text{s.t.}: & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i, \ \ i = 1, \cdots, m, \\ & \|\mathbf{w}\|^* \leq \alpha, \\ & \xi_i \geq 0, \ \ i = 1, \cdots, m. \end{array} \tag{16}$$

We notice that Problem (16) is the standard regularized classifier. Hence, the comprehensive robust classification framework is a general framework which includes both robust SVMs and regularized SVMs as special cases. Hence, the results obtained for the comprehensive robust classifier (e.g., the probabilistic bound in Section 3.4) can be easily applied to robust SVMs and standard SVMs.

3.3 Tractability

We now give a sufficient condition on the discount for the comprehensive robust classification problem (15) to be tractable.

Definition 8 *A function $f(\cdot) : \mathbb{R}^n \to \mathbb{R}$ is called* Efficiently Conjugatable *if there exists a sub-routine such that for arbitrary $\mathbf{h} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, in polynomial time it either reports*

$$\sup_{\mathbf{x} \in \mathbb{R}^n} \left(\mathbf{h}^\top \mathbf{x} - f(\mathbf{x}) \right) \leq \alpha,$$

or reports $\mathbf{x}_0$ such that

$$\mathbf{h}^\top \mathbf{x}_0 - f(\mathbf{x}_0) > \alpha.$$

Theorem 9 *Suppose*

1. *$f_i(\cdot)$ is efficiently conjugatable, $\forall i \in [1 : m]$.*
2. *Both $r(\mathbf{w}, b)$ and $\partial r(\mathbf{w}, b)$ can be evaluated in polynomial time $\forall (\mathbf{w}, b) \in \mathbb{R}^{n+1}$, where ∂ stands for any sub-gradient.*

Then, Problem (15) can be solved in polynomial time.

Proof Rewrite Problem (15) as

$$\begin{array}{ll} \min: & t \\ \text{s.t.}: & r(\mathbf{w}, b) + \sum_{i=1}^m \xi_i - t \leq 0 \\ & f_i^*(y_i\mathbf{w}) - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - \xi_i + 1 \leq 0, \ \ i = 1, \cdots, m, \\ & -\xi_i \leq 0, \ \ i = 1, \cdots, m. \end{array} \tag{17}$$

This is a special case of $\min_{\mathbf{z}\in\mathcal{U}} \mathbf{c}^\top \mathbf{z}$ for a convex $\mathcal{U}$. It is known (Grótschel et al., 1988) that for this problem to be efficiently solvable, it suffices to have a *Separation Oracle* for $\mathcal{U}$, i.e., a subroutine which in polynomial time reports either $\mathbf{z} \in \mathcal{U}$, or a separating hyperplane of $\mathbf{z}$ and $\mathcal{U}$ when $\mathbf{z} \notin \mathcal{U}$ for any $\mathbf{z}$.

We can construct a separation oracle for $\mathcal{U}$ as long as we can construct a separation oracle for the feasible set of each individual constraint.

Constraint Type 1: $r(\mathbf{w}, b) + \sum_{i=1}^m \xi_i - t \leq 0$.

For any $(\mathbf{w}^*, \boldsymbol{\xi}^*, t^*, b^*)$, since $r(\mathbf{w}^*, b^*)$ can be evaluated efficiently, we can report whether this constraint holds or not in polynomial time. Furthermore, when the constraint is violated, any sub-gradient of the left-hand side evaluated at $(\mathbf{w}^*, \boldsymbol{\xi}^*, t^*, b^*)$ is a separating hyperplane. Finding such sub-gradient can also be done efficiently since $\partial r(\mathbf{w}^*, b)$ can be evaluated efficiently.

Constraint Type 2: $f_i^*(y_i \mathbf{w}) - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - \xi_i + 1 \leq 0$.

For given $(\mathbf{w}^*, \boldsymbol{\xi}^*, t^*, b^*)$, let $\alpha = y_i(\langle \mathbf{w}^*, \mathbf{x}_i \rangle + b^*) + \xi_i^* - 1$, and $\mathbf{h} = y_i \mathbf{w}^*$. Since $f_i(\cdot)$ is efficiently conjugatable, in polynomial time we can either confirm

$$\sup_{\mathbf{c}\in\mathbb{R}^n} \left(\mathbf{h}^\top \mathbf{c} - f(\mathbf{c})\right) \leq \alpha,$$

which means the constraint holds, or report a $\mathbf{c}_0$ such that

$$\mathbf{h}^\top \mathbf{c_0} - f(\mathbf{c_0}) > \alpha.$$

Substituting back α, $\mathbf{h}$ and rearranging the terms yields

$$(y_i \mathbf{c}_0 - y_i \mathbf{x}_i)^\top \mathbf{w}^* - y_i b^* - \xi_i^* > f(\mathbf{c}_0) - 1.$$

Notice that, for any feasible $(\hat{\mathbf{w}}, \hat{\boldsymbol{\xi}}, \hat{t}, \hat{b})$, the following holds:

$$\begin{aligned}
& f_i^*(y_i \hat{\mathbf{w}}) - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) - \hat{\xi}_i + 1 \leq 0 \\
\Longrightarrow \quad & \sup_{\mathbf{c}\in\mathbb{R}^n} \left(y_i \hat{\mathbf{w}}^\top \mathbf{c} - f(\mathbf{c})\right) - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) - \hat{\xi}_i + 1 \leq 0 \\
\Longrightarrow \quad & \left(y_i \hat{\mathbf{w}}^\top \mathbf{c}_0 - f(\mathbf{c}_0)\right) - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) - \hat{\xi}_i + 1 \leq 0 \\
\Longrightarrow \quad & (y_i \mathbf{c}_0 - y_i \mathbf{x}_i)^\top \hat{\mathbf{w}} - y_i \hat{b} - \hat{\xi}_i \leq f(\mathbf{c}_0) - 1.
\end{aligned}$$

Hence $(y_i \mathbf{c}_0 - y_i \mathbf{x}_i, -y_i, -1)$ is a separation Oracle.

Constraint Type 3: $-\xi_i \leq 0$.

The separation oracle for this constraint is trivial.

Combining all three steps, we conclude that a separation oracle exists for each individual constraint, and hence we have a separation Oracle for $\mathcal{U}$. Therefore, Problem (15) can be solved in polynomial time. ■

This theorem guarantees polynomial time solvability, but much stronger complexity requirements may be needed for large scale problems. While this is a topic of future research, in the nest section we provide some discount function examples that are of practical interest.

3.4 Norm Discount

In this section, we discuss a class of discount functions based on certain ellipsoidal norms of the noise, i.e.,

$$f_i(\boldsymbol{\delta}_i) = t_i(\|\boldsymbol{\delta}\|_V),$$

for a nondecreasing $t_i : \mathbb{R}^+ \to \mathbb{R}^+$. Simple algebra yields $f_i^*(\mathbf{y}) = t_i^*(\|\mathbf{y}\|_{V^{-1}})$, where $t_i^*(y) = \sup_{x\geq 0}\big[xy - t(x)\big]$, and thus conjugation is easy. This formulation has a nice probabilistic interpretation:

Theorem 10 *Suppose the random variable $\boldsymbol{\delta}_i^r$ has mean $\mathbf{0}$ and variance Σ. Then the constraint*

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - t_i(\|\boldsymbol{\delta}_i\|_{\Sigma^{-1}}),\ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n, \tag{18}$$

is equivalent to

$$\inf_{\boldsymbol{\delta}_i^r \sim (0,\Sigma)} Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) \geq 1 - \frac{1}{\big(t_i^{-1}(s)\big)^2 + 1},\ \forall s \geq 0. \tag{19}$$

Here, the infimum is taken over all random variables with mean zero and variance Σ, and $t_i^{-1}(s) \triangleq \sup\{r | t(r) \leq x\}$.

Proof Shivaswamy et al. (2006) studied the robust formulation and showed that for a fixed γ_0, the following three inequalities are equivalent:

- $\inf_{\boldsymbol{\delta}_i^r \sim (0,\Sigma)} Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq 0\big) \geq 1 - \frac{1}{\gamma_0^2 + 1}$,
- $y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - 1 + \xi_i \geq \gamma_0 \|\mathbf{w}\|_\Sigma$,
- $y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) - 1 + \xi_i \geq 0, \quad \forall \|\boldsymbol{\delta}_i\|_{\Sigma^{-1}} \leq \gamma_0$.

Observe that equation (19) is equivalent to

$$\inf_{\boldsymbol{\delta}_i^r \sim (0,\Sigma)} Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -t_i(\gamma)\big) \geq 1 - \frac{1}{\gamma^2 + 1},\ \forall \gamma \geq 0.$$

Hence, it is equivalent to:

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) - 1 + \xi_i \geq -t_i(\gamma),\ \forall \|\boldsymbol{\delta}_i\|_{\Sigma^{-1}} \leq \gamma, \quad \forall \gamma \geq 0.$$

Since $t_i(\cdot)$ is nondecreasing, this is equivalent to (18). ■

Theorem 10 shows that the comprehensive robust formulation bounds the probability of *all* magnitudes of constraint violation. It is of interest to compare this bound with the bound given by the robust formulation. Indeed,

$$\begin{aligned}
& y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - 1 + \xi_i \geq \gamma_0 \|\mathbf{w}\|_\Sigma \\
\iff \quad & y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - 1 + \xi_i + s \geq \big(\gamma_0 + \frac{s}{\|\mathbf{w}\|_\Sigma}\big)\|\mathbf{w}\|_\Sigma,\ \forall s \geq 0 \\
\iff \quad & \inf_{\boldsymbol{\delta}_i^r \sim (0,\,\Sigma)} Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) \geq 1 - \frac{1}{(\gamma_0 + \frac{s}{\|\mathbf{w}\|_\Sigma})^2 + 1}.
\end{aligned}$$

Hence the probability of large violation depends on $\|\mathbf{w}\|_\Sigma$, and is impossible to bound without knowing $\|\mathbf{w}\|_\Sigma$ *a priori*.

Remark 11 Notice the derived bound for the robust formulation is tight, in the sense that if

$$y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1 + \xi_i < \gamma_0 \|\mathbf{w}\|_\Sigma,$$

then there exists a zero-mean random variable $\boldsymbol{\delta}_i^r$ with variance Σ such that

$$Pr\big(y_i(\langle \mathbf{w}, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) < 1 - \frac{1}{(\gamma_0 + \frac{s}{\|\mathbf{w}\|_\Sigma})^2 + 1}.$$

This is because the multivariate Chebyshev inequality (Marshall and Olkin, 1960; Boyd and Vandenberghe, 2004; Grótschel et al., 1988) states that

$$\begin{aligned} &\sup_{\mathbf{z}\sim(\bar{\mathbf{z}},\sigma)} Pr\{\mathbf{a}^\top \mathbf{z} \leq c\} = (1+d^2)^{-1} \\ \text{where} \qquad &d^2 = \inf_{\mathbf{z}_0 | \mathbf{a}^\top \mathbf{z}_0 \leq c} \inf (\mathbf{z}_0 - \bar{\mathbf{z}})^\top \Sigma^{-1} (\mathbf{z}_0 - \bar{\mathbf{z}}). \end{aligned}$$

Letting $\mathbf{a} = y_i \mathbf{w}$, $\mathbf{z} = -\boldsymbol{\delta}_i^r$ and $c = 1 - \xi_i - s - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b)$, we have

$$\begin{aligned} &\sup_{\boldsymbol{\delta}_i^r \sim (0,\Sigma)} Pr\big(y_i(\langle \mathbf{w}, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \leq -s\big) = (1+d_0^2)^{-1} \\ \text{where:} \quad &d_0 = \frac{y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1 + \xi_i + s}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}}. \end{aligned}$$

Hence,

$$\begin{aligned} & y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1 + \xi_i < \gamma_0 \|\mathbf{w}\|_\Sigma \\ \Longrightarrow \quad & d_0 < \gamma_0 + s/\|\mathbf{w}\|_\Sigma \\ \Longrightarrow \quad & \sup_{\boldsymbol{\delta}_i^r \sim (0,\Sigma)} Pr\big(y_i(\langle \mathbf{w}, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \leq -s\big) > \big[1 + (\gamma_0 + s/\|\mathbf{w}\|_\Sigma)^2\big]^{-1}, \end{aligned}$$

showing that the bound is tight.

With a similar argument, we can derive probability bounds under a Gaussian noise assumption.

Theorem 12 *If $\boldsymbol{\delta}_i^r \sim \mathcal{N}(0,\Sigma)$, then the constraint*

$$y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - t_i(\|\boldsymbol{\delta}_i\|_{\Sigma^{-1}}),\ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n, \tag{20}$$

is equivalent to

$$Pr\big(y_i(\langle \mathbf{w}, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) \geq \Phi\big(t_i^{-1}(s)\big), \qquad \forall s \geq 0. \tag{21}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$.

Proof For fixed $k \geq 1/2$ and constant l, the following constraints are equivalent:

$$\begin{aligned} & Pr(y_i \mathbf{w}^\top \boldsymbol{\delta}_i^r \geq l) \geq k \\ \Longleftrightarrow \quad & l \leq \Phi^{-1}(k)\big(\mathbf{w}^\top \Sigma \mathbf{w}\big)^{1/2} \\ \Longleftrightarrow \quad & l \leq y \mathbf{w}^\top \boldsymbol{\delta}_i,\ \forall \|\boldsymbol{\delta}_i\|_{\Sigma^{-1}} \leq \Phi^{-1}(k). \end{aligned}$$

	Original Function	Conjugate function
Affine Fun.	$t_i(x) = ax + b$	$t_i^*(y) = I_\alpha - b$
Indicate Fun.	$t_i(x) = I_\alpha + b.$	$t_i^*(y) = ay - b$
Power Fun.	$t_i(x) = ax^n + b$	$t_i^*(y) = a^{\frac{-1}{n-1}}(n^{\frac{-1}{n-1}} - n^{\frac{-n}{n-1}})y^{\frac{n}{n-1}} - b$
Quadratic Fun.	$t_i(x) = ax^2 + b$	$t_i^*(y) = \frac{1}{4a}y^2 - b$
Neg. Entropy	$t_i(x) = ax\log x + b$	$t_i^*(y) = ae^{y/a-1} - b$
Exponential Fun.	$t_i(x) = ae^x + b$	$t_i^*(y) = y\log(y/a) - y - b$
Point-wise Min.	$t_i(x) = \min_{j=1,\cdots,l} t_{ij}(x)$	$t_i^*(x) = \max_{j=1,\cdots,l} t_{ij}^*(y)$
Non-convex Fun.	$t_i(x)$ non-convex	$t_i^*(y) = \big(conv(t_i(x))\big)^*(y)$

Table 1: Some functions and their conjugates.

Notice that (21) is equivalent to

$$Pr\Big(y_i(\langle \mathbf{w}, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -t_i(\gamma)\Big) \geq \Phi(\gamma),\ \forall \gamma \geq 0,$$

and hence it is equivalent to: $\forall \gamma \geq 0$,

$$y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) - 1 + \xi_i \geq -t_i(\gamma), \quad \forall \|\boldsymbol{\delta}_i\|_{\Sigma^{-1}} \leq \Phi^{-1}\big(\Phi(\gamma)\big) = \gamma.$$

Since $t_i(\cdot)$ is nondecreasing, this is equivalent to (20). ■

We list in Tables 1 and 2 some examples of $t_i(\cdot)$ and their conjugate functions. Notice that both $t_i(\cdot)$ and $t_i^*(\cdot)$ are defined on $\mathbb{R}^+$. Here, $I_\alpha : \mathbb{R}^+ \to \mathbb{R}^+ \bigcup\{+\infty\}$ is the indicator function of set α, and $conv(t(\cdot)) \triangleq \sup\{f(\cdot)|f(\cdot)$ is convex, $f(\cdot) \leq t(\cdot)\}$, i.e., the convex supporting function of $t(\cdot)$. Standard robustness uses an indicator function of a set. Table 2 shows several different relaxations of this indicator function allowing the increase of $f(\cdot)$ to be more smooth. Notice that, all conjugate functions can be written as $t^*(x) = \max_{1,2}(s_1(x), s_2(x))$, where $s_i = \inf_{\boldsymbol{\lambda}\in\mathcal{S}_i} q_i(\lambda, x)$ for some quadratic/linear/indicator function q_i and polytope $\mathcal{S}_i$. Hence the constraint $t_i^*(x) \leq \alpha$ is equivalent to

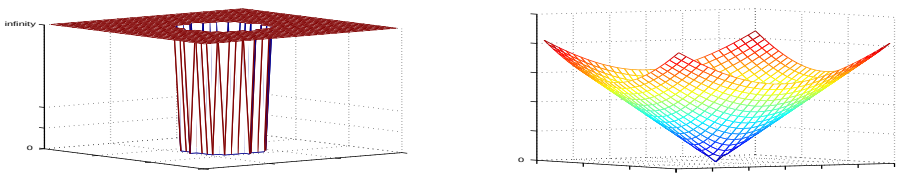
$$\begin{aligned} &q_1(x, \boldsymbol{\lambda}_1) \leq \alpha; \\ &\boldsymbol{\lambda}_1 \in \mathcal{S}_1; \\ &q_2(x, \boldsymbol{\lambda}_2) \leq \alpha; \\ &\boldsymbol{\lambda}_2 \in \mathcal{S}_2. \end{aligned}$$

Since a quadratic/linear/indicator function leads to a Second Order Cone constraint, the resulting classifier is a SOCP. This means that the comprehensive robust classification with the relaxations listed above has a comparable computational cost as the robust classification.

Figure 1 illustrates the discount function for the standard robust formulation, and Figure 2 illustrates the respective conjugate functions for the first four relaxations in Table 2.

Original Function	Conjugate function
$t_i(x) = \begin{cases} 0 & x \le c, \\ \alpha(x-c) & x > c. \end{cases}$	$t_i^*(y) = \begin{cases} cy & y \le \alpha, \\ +\infty & y > \alpha. \end{cases}$ $= \max(I_\alpha,\, cy)$
$t_i(x) = \begin{cases} \alpha x & x \le c, \\ +\infty & x > c. \end{cases}$	$t_i^*(y) = \begin{cases} 0 & y \le \alpha, \\ c(y-\alpha) & y > \alpha. \end{cases}$ $= \max(0,\, c(y-\alpha))$
$t_i(x) = \begin{cases} 0 & x \le c_1, \\ \alpha(x-c_1) & c_1 < x \le c_2, \\ +\infty & x > c_2. \end{cases}$	$t_i^*(y) = \begin{cases} c_1 y & y \le \alpha, \\ c_2(y-\alpha)+\alpha c_1 & y > \alpha. \end{cases}$ $= \max(c_1 y,\, c_2 y + \alpha(c_1 - c_2))$
$t_i(x) = \begin{cases} 0 & x \le c, \\ \alpha(x-c)^2 & x > c. \end{cases}$	$t_i^*(y) = y^2/4\alpha + cy.$
$t_i(x) = \begin{cases} \alpha x^2 & x \le c, \\ +\infty & x > c. \end{cases}$	$t_i^*(y) = \begin{cases} y^2/4\alpha & y \le 2\alpha c, \\ cy - \alpha c^2 & y > 2\alpha c. \end{cases}$ $= \inf_{\lambda \ge 0}\left((y-\lambda)^2/4\alpha + c\lambda\right)$
$t_i(x) = \begin{cases} 0 & x \le c_1, \\ \alpha(x-c_1)^2 & c_1 < x \le c_2, \\ +\infty & x > c_2. \end{cases}$	$t_i^*(y) = \begin{cases} y^2/4\alpha + yc_1 & y \le 2\alpha(c_2-c_1), \\ c_2 y - \alpha(c_2-c_1)^2 & y > 2\alpha(c_2-c_1). \end{cases}$ $= \max\left(c_1 y,\, \inf_{\lambda_1,\lambda_2 \ge 0}[\frac{(y+\lambda_1-\lambda_2)^2}{4\alpha} + c_1 y + (c_2-c_1)\lambda_2]\right)$

Table 2: Piecewise-defined functions and their conjugates.



(r) (r')

Figure 1: The robust discount function and its conjugate: note that the discount function provides uniform protection inside the uncertainty set, and no protection outside.

3.5 Multiplicative Discount

In this section we consider a multiplicative structure for the disturbance discount, and investigate its tractability. The multiplicative discount has the form:

$$\min_{\mathbf{w},b} \max_{(\boldsymbol{\delta}_1,\cdots\boldsymbol{\delta}_m)\in\mathcal{N}} \left\{ r(\mathbf{w},b) + \sum_{i=1}^{m} c_i(\boldsymbol{\delta}_i) \max\left[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b),\, 0\right] \right\}$$

where $c(\cdot) : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$0 \le c_i(\boldsymbol{\delta}) \le c_i(\mathbf{0}) = 1; \quad \forall \boldsymbol{\delta} \in \mathbb{R}^n.$$

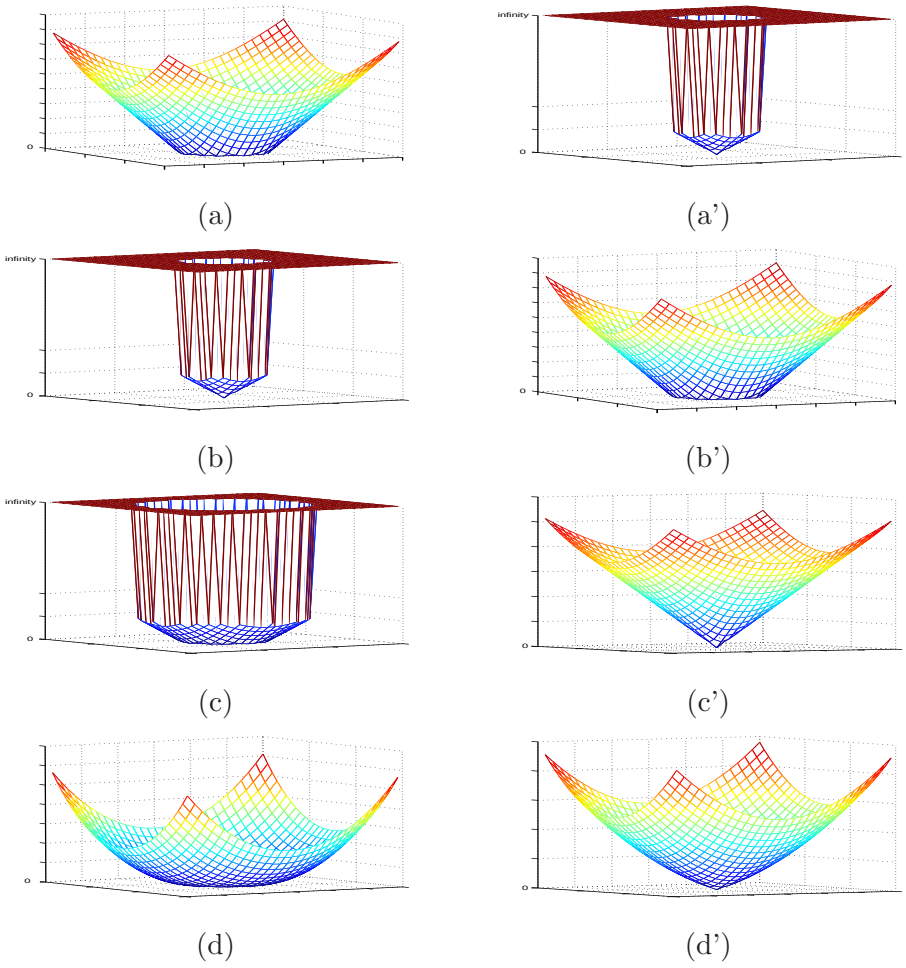


Figure 2: Piecewise-defined Functions and their Conjugates: note the flexibility in controlling the discount given the realization of the disturbance.

By adding slack variables, we get the following optimization problem:

$$\begin{array}{ll} & \textbf{Comprehensive Robust Classifier (Multiplicative):} \\ \min: & r(\mathbf{w}, b) + \sum_{i=1}^{m} \xi_i, \\ \text{s.t.}: & \xi_i \geq c_i(\boldsymbol{\delta}) \big[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b)\big], \qquad \forall \boldsymbol{\delta}_i \in \mathbb{R}^n,\ i = 1, \cdots, m, \\ & \xi_i \geq 0, \quad i = 1, \cdots, m. \end{array} \tag{22}$$

Define

$$g_i(\boldsymbol{\delta}) \triangleq \begin{cases} \frac{1}{c_i(\boldsymbol{\delta})} & \text{if } c(\boldsymbol{\delta}) > 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Problem (22) can be rewritten as:

$$\begin{array}{rl} \min: & r(\mathbf{w},b)+\sum_{i=1}^m \xi_i, \\ \text{s.t.}: & g_i(\boldsymbol{\delta}_i)\xi_i \geq \left[1-y_i(\langle \mathbf{w},\, \mathbf{x}_i-\boldsymbol{\delta}_i\rangle + b)\right], \qquad \forall \boldsymbol{\delta}_i \in \mathbb{R}^n,\ i=1,\cdots,m, \\ & \xi_i \geq \epsilon, \quad i=1,\cdots,m. \end{array}$$

We perturb the constraint $\xi_i \geq 0$ to $\xi_i \geq \epsilon$ for small $\epsilon > 0$ to avoid the case that both $\xi_i = 0$ and $g_i(\boldsymbol{\delta}_i) = \infty$ hold simultaneously. Under this modification, we have the following tractability theorem:

Theorem 13 *Suppose*

1. *$g_i(\cdot)$ is efficiently conjugatable, $\forall i \in [1:m]$*
2. *Both $r(\mathbf{w},b)$, $\partial r(\mathbf{w},b)$ can be evaluated in polynomial time $\forall (\mathbf{w},b) \in \mathbb{R}^{n+1}$, where ∂ stands for any sub-gradient.*

Then, Problem (22) can be solved in polynomial time.

Proof Rewrite Problem (22) as

$$\begin{array}{rl} \min: & t \\ \text{s.t.}: & r(\mathbf{w},b)+\sum_{i=1}^m \xi_i - t \leq 0 \\ & 1-y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) + y_i \mathbf{w}^\top \boldsymbol{\delta}_i - \xi_i g_i(\boldsymbol{\delta}_i) \leq 0, \qquad \forall \boldsymbol{\delta}_i \in \mathbb{R}^n,\ i=1,\cdots,m, \\ & -\xi_i \leq -\epsilon, \quad i=1,\cdots,m. \end{array}$$

Following a similar argument as in the proof of Theorem 9 we derive a separation oracle for each constraint. Constraint Type 1 and Type 3 are exactly the same as in Theorem 9, hence we only discuss Constraint Type 2, i.e.,

$$1-y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) + y_i \mathbf{w}^\top \boldsymbol{\delta}_i - \xi_i g_i(\boldsymbol{\delta}_i) \leq 0,\ \forall \boldsymbol{\delta}_i.$$

Now suppose we are given a solution $(\mathbf{w}^*, \boldsymbol{\xi}^*, t^*, b^*)$, with $\xi_i^* \geq \epsilon$ (otherwise we get a separation oracle from Type 3). Letting $\mathbf{h} = y_i \mathbf{w}^*/\xi_i^*$ and $\alpha = (1-y_i(\langle \mathbf{w}^*,\, \mathbf{x}_i\rangle + b^*))/\xi_i^*$, the constraint is equivalent to:

$$\sup_{\boldsymbol{\delta}_i \in \mathbb{R}^n} \left\{ \mathbf{h}^\top \boldsymbol{\delta}_i - g_i(\boldsymbol{\delta}_i) \right\} \leq \alpha.$$

Since $g_i(\cdot)$ is efficiently conjugatable, then in polynomial time we either conclude the constraint is satisfied, or find a $\boldsymbol{\delta}^*$ such that $\mathbf{h}^\top \boldsymbol{\delta}^* - g_i(\boldsymbol{\delta}^*) > \alpha$, which is equivalent to

$$\begin{array}{rl} & 1-y_i(\langle \mathbf{w}^*,\, \mathbf{x}_i\rangle + b^*) + y_i \mathbf{w}^{*\top} \boldsymbol{\delta}^* - \xi_i^* g_i(\boldsymbol{\delta}^*) > 0 \\ \iff & y(\mathbf{x}_i - \boldsymbol{\delta}^*)^\top \mathbf{w}^* + y_i b^* + g_i(\boldsymbol{\delta}^*)\xi_i < 1. \end{array} \tag{23}$$

Hence, $(\mathbf{x}_i - \boldsymbol{\delta}^*, y_i, g_i(\boldsymbol{\delta}^*))$ is a separation oracle. ■

4. Comprehensive Robustness and Convex Risk Measures

We showed in Section 2.2 that the robust optimization classifier has an equivalent probabilistic interpretation as a chance constrained classifier. Comprehensive robust classifiers under the additive discount model also have a probabilistic parallel. In this section we establish the connection to risk-measure constrained classifiers. A risk measure is a mapping from a random variable to the real numbers, that, at a high level, captures some valuation of that random variable. Simple examples of risk measures include expectation, standard deviation, and conditional value at risk (CVaR). Risk measure constraints represent a natural way to express risk aversion, corresponding to particular risk preferences. We show that comprehensive robust classifiers correspond to the class of so-called convex risk measures.

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{X}$ denote the set of random variables on Ω. A *risk measure* is a function $\rho : \mathcal{X} \to \mathbb{R}$, and defines a preference relationship among random variables: X_1 is preferable over X_2 if and only if $\rho(X_1) \leq \rho(X_2)$. Alternatively, we can regard $\rho(\cdot)$ as the measurement of how risky a random variable is: X_1 is a less risky decision than X_2 when $\rho(X_1) \leq \rho(X_2)$. A risk measure is called *convex* if it satisfies the following three conditions: (i) Convexity: $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$; (ii) Monotonicity: $X \leq Y \Rightarrow \rho(X) \leq \rho(Y)$; and (iii) Translation Invariance: $\rho(X + a) = \rho(X) + a, \forall a \in \mathbb{R}$. Convexity means diversifying reduces risk. Monotonicity says that if one random loss is always less than another, it is more favorable. Translation invariance says that if a fixed penalty a is going to be paid in addition to X, we are indifferent to whether we will pay it before or after X is realized. A convex risk measure $\rho(\cdot)$ is called *normalized* if it satisfies $\rho(0) = 0$ and $\forall X \in \mathcal{X}$, $\rho(X) \geq \mathbb{E}_{\mathbb{P}}(X)$, which essentially says that the risk measure $\rho(\cdot)$ represents risk aversion. Many widely used criteria comparing random variables are normalized convex risk measures, including expected value, Conditional Value at Risk (CVaR), and the exponential loss function (Ben-Tal et al., 2006b; Bertsimas and Brown, 2005).

Equipped with a normalized convex risk measure $\rho(\cdot)$, we can formulate a classification problem as follows:

Risk-Measure Constrained Classifier

$$\begin{aligned} \min: \quad & r(\mathbf{w}, b) + \sum_{i=1}^{m} \xi_i, \\ \text{s.t.}: \quad & \rho_i(\xi_i) \geq \rho_i(1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i^r \rangle + b)), \;\; i = 1, \cdots, m, \\ & \xi_i \geq 0, \;\; i = 1, \cdots, m. \end{aligned} \tag{24}$$

Substituting $\rho_i(0) = 0$ and $\mathbf{x}_i^r = \mathbf{x}_i - \boldsymbol{\delta}_i^r$ where $\mathbf{x}_i = \mathbb{E}_{\mathbb{P}}(\mathbf{x}_i^r)$, the constraint can be rewritten as

$$\xi_i \geq 1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i \rangle + b) + \rho_i(y_i \mathbf{w}^\top \boldsymbol{\delta}_i^r). \tag{25}$$

This formulation seeks a classifier whose total risk is minimized. When $\mathbf{x}_i^r$ is precisely known, this formulation reduces to the standard SVM.

The following theorem states that the risk-constrained classifier and the comprehensive robust classifier are equivalent. The proof is postponed to the appendix.

Theorem 14 *(1) A Risk-Measure Constrained Classifier with normalized convex risk measures $\rho_i(\cdot)$ is equivalent to a Comprehensive Robust Classifier where*

$$\begin{aligned} f_i(\boldsymbol{\delta}) &= \inf\{\alpha_i^0(Q) | \mathbb{E}_Q(\boldsymbol{\delta}_i^r) = \boldsymbol{\delta}\}, \\ \alpha_i^0(Q) &\triangleq \sup_{X' \in \mathcal{X}} \big(\mathbb{E}_Q(X') - \rho_i(X')\big). \end{aligned}$$

(2) A Comprehensive Robust Classifier with convex discount functions $f_i(\cdot)$ is equivalent to a Risk-Constrained Classifier where

$$\begin{aligned} \rho_i(X) &= \inf\{m \in \mathbb{R} | X - m \in \mathcal{A}_i\}, \\ \mathcal{A}_i &\triangleq \{X \in \mathcal{X} | X(\omega) \leq f_i\big(\boldsymbol{\delta}_i^r(\omega)\big),\ \forall \omega \in \Omega\}, \end{aligned}$$

assuming that $\boldsymbol{\delta}_i^r$ has support $\mathbb{R}^n$.

Let $\mathcal{P}$ be the set of probability measures absolutely continuous w.r.t. $\mathbb{P}$. It is known (Fóllmer and Schied, 2002; Ben-Tal et al., 2006a) that any convex risk measure $\rho(\cdot)$ can be represented as $\rho(X) = \sum_{Q \in \mathcal{P}} [\mathbb{E}_Q(X) - \alpha(Q)]$ for some convex function $\alpha(\cdot)$; conversely, given any such convex function α, the resulting function $\rho(\cdot)$ is indeed a convex risk measure. Given $\alpha(\cdot)$, $\rho(\cdot)$ is called the corresponding risk measure. The function $\alpha(\cdot)$ can be thought of as a penalty function on probability distributions. This gives us a way to directly investigate classifier robustness with respect to distributional deviation. As an example, suppose we want to be robust over distributions that are nowhere more than a factor of two greater than a nominal distribution, $\mathbb{P}$. This can be exactly captured by the risk constraint using risk measure $\rho(\cdot)$, where ρ corresponds to the convex function α given by letting $\alpha(\cdot)$ satisfy $\alpha(Q) = 0$ for $dQ/d\mathbb{P} \leq 2$, and $\alpha(Q) = +\infty$ for all other Q.

A natural notion of distributional divergence is the Kullback-Leibler divergence. The next result derives the corresponding risk measure when the reference noise, $\boldsymbol{\delta}_i^r$, is Gaussian.

Theorem 15 *Suppose $\boldsymbol{\delta}_i^r \sim \mathcal{N}(0, \Sigma_i)$ and let $\rho(\cdot)$ be the corresponding risk measure of*

$$\alpha(Q) = \begin{cases} \int \frac{dQ}{d\mathbb{P}} \log \frac{dQ}{d\mathbb{P}} d\mathbb{P} & Q \ll \mathbb{P}, \\ +\infty & \textit{otherwise.} \end{cases}$$

Then the Risk-Measure Constrained Classifier is equivalent to

$$\begin{aligned} \min: \quad & r(\mathbf{w}, b) + \sum_{i=1}^m \xi_i, \\ \text{s.t.}: \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - \mathbf{w}^\top \Sigma_i \mathbf{w}/2 \geq 1 - \xi_i, \ \ i = 1, \cdots, m, \\ & \xi_i \geq 0, \ \ i = 1, \cdots, m. \end{aligned}$$

Proof We first show that for the KL divergence, its corresponding convex risk measure equals $\log \mathbb{E}_{\mathbb{P}}[e^X]$ by applying the following theorem adapted from Fóllmer and Schied (2002).

Theorem 16 *Suppose a convex risk measure can be represented as*

$$\rho(X) = \inf\{m \in \mathbb{R} | \mathbb{E}_{\mathbb{P}}[l(X - m)] \leq x_0\},$$

for an increasing convex function $l : \mathbb{R} \to \mathbb{R}$ and scalar x_0. Then $\rho(\cdot)$ is the corresponding risk measure of

$$\alpha_0(Q) = \inf_{\lambda>0} \frac{1}{\lambda}\left(x_0 + \mathbb{E}_{\mathbb{P}}\big[l^*(\lambda\frac{dQ}{d\mathbb{P}})\big]\right).$$

Note that $\log \mathbb{E}_{\mathbb{P}}[e^X] = \inf\{m \in \mathbb{R} | \mathbb{E}_{\mathbb{P}}\big[e^{X-m}\big] \leq 1\}$, and hence the risk measure $\log \mathbb{E}_{\mathbb{P}}[e^X]$ can be represented as in the theorem, with $l(x) = e^x$, and $x_0 = 1$. The conclusion of the theorem tells us that $\log \mathbb{E}_{\mathbb{P}}[e^X]$ is the corresponding risk measure of

$$\begin{aligned}
&\alpha_0(Q) = \inf_{\lambda>0} \frac{1}{\lambda}\left(1 + \mathbb{E}_{\mathbb{P}}\big[\lambda\frac{dQ}{d\mathbb{P}}\log(\lambda\frac{dQ}{d\mathbb{P}}) - \lambda\frac{dQ}{d\mathbb{P}}\big]\right)\\
=&\mathbb{E}_{\mathbb{P}}\big[\frac{dQ}{d\mathbb{P}}\log\frac{dQ}{d\mathbb{P}}\big] + \inf_{\lambda>0}\big[\frac{1}{\lambda} + \mathbb{E}_{\mathbb{P}}(\frac{dQ}{d\mathbb{P}})(\log\lambda - 1)\big]\\
=&\begin{cases} \int \frac{dQ}{d\mathbb{P}}\log\frac{dQ}{d\mathbb{P}}d\mathbb{P} & Q \ll \mathbb{P},\\ +\infty & \text{otherwise,}\end{cases}
\end{aligned}$$

where the last equation holds since $\mathbb{E}_{\mathbb{P}}(dQ/d\mathbb{P}) = 1$ and $\inf_{\lambda>0}(1/\lambda + \log\lambda - 1) = 0$. Therefore $\rho(X) = \log \mathbb{E}_{\mathbb{P}}[e^X]$ is indeed the corresponding risk measure to KL-divergence. Now we evaluate $\log \mathbb{E}_{\mathbb{P}}(e^{y_i\mathbf{w}^\top\boldsymbol{\delta}_i^r})$. Since $\boldsymbol{\delta}_i^r \sim N(0, \Sigma_i)$, $y_i\mathbf{w}^\top\boldsymbol{\delta}_i^r \sim N(0, \mathbf{w}^\top\Sigma_i\mathbf{w})$, which leads to

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}}(e^{y_i\mathbf{w}^\top\boldsymbol{\delta}_i^r}) = \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\Big[-t^2/2\sqrt{\mathbf{w}^\top\Sigma_i\mathbf{w}}\Big]e^t dt\\
&= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\Big\{-(t-\sqrt{\mathbf{w}^\top\Sigma_i\mathbf{w}})^2\Big/2\sqrt{\mathbf{w}^\top\Sigma_i\mathbf{w}}\Big\}e^{\mathbf{w}^\top\Sigma_i\mathbf{w}/2}dt\\
&= e^{\mathbf{w}^\top\Sigma_i\mathbf{w}/2}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\Big\{-(t-\sqrt{\mathbf{w}^\top\Sigma_i\mathbf{w}})^2\Big/2\sqrt{\mathbf{w}^\top\Sigma_i\mathbf{w}}\Big\}dt = e^{\mathbf{w}^\top\Sigma_i\mathbf{w}/2}.
\end{aligned}$$

Thus $\log \mathbb{E}_{\mathbb{P}}(e^{y_i\mathbf{w}^\top\boldsymbol{\delta}_i^r}) = \mathbf{w}^\top\Sigma_i\mathbf{w}/2$, proving the theorem. ■

Observe that here we get a regularizer (in each constraint) that is the *square* of an ellipsoidal norm, and hence is different from the norm regularizer obtained from the robust classification framework. In fact, recalling the result from Section 3.4, we notice that the new regularizer is the result of a quadratic discount function, instead of the indicator discount function used by robust classification.

For general $\boldsymbol{\delta}_i^r$ and $\alpha(\cdot)$, it is not always straightforward to find and optimize the explicit form of the regularization term. Hence we sample, approximating $\mathbb{P}$ with its empirical distribution $\mathbb{P}_n$. This is equivalent to assuming $\boldsymbol{\delta}_i^r$ has finite support $\{\boldsymbol{\delta}_i^1, \cdots, \boldsymbol{\delta}_i^t\}$ with probability $\{p_1, \cdots, p_t\}$. We note that the distribution of the noise is often unknown, where only some samples of the noise are given. Therefore, the finite-support approach is often an appropriate method in practice.

Theorem 17 *For $\boldsymbol{\delta}_i^r$ with finite support, the risk-measure constrained classifier is equivalent to*

$$\begin{aligned}\min: \quad & r(\mathbf{w},b)+\sum_{i=1}^{m}\xi_i,\\ \text{s.t.}: \quad & y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - \alpha^*\big(y_i\Delta_i^\top \mathbf{w}+\lambda_i\mathbf{1}\big)+\lambda_i \geq 1-\xi_i,\ i=1,\cdots,m;\\ & \xi_i\geq 0,\ i=1,\cdots,m;\end{aligned}$$

where $\alpha^*(\mathbf{y}) \triangleq \sup_{\mathbf{x}\geq\mathbf{0}}\{\mathbf{y}^\top\mathbf{x}-\alpha(\mathbf{x})\}$ and $\Delta_i \triangleq \{\boldsymbol{\delta}_i^1,\cdots,\boldsymbol{\delta}_i^t\}$.

Proof It suffices to prove that Constraint (25) is equivalent to

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - \alpha^*\big(y_i\Delta_i^\top \mathbf{w}+\lambda_i\mathbf{1}\big)+\lambda_i \geq 1-\xi_i,$$

which is the same as showing that the conjugate function of

$$f_i(\boldsymbol{\delta}) \triangleq \inf\{\alpha(\mathbf{q})|\sum_{j=1}^{t} q_j\delta_i^j=\boldsymbol{\delta}\}$$

evaluated at $y_i\mathbf{w}$ equals

$$\min_{\lambda}\{\alpha^*\big(y_i\Delta_i^\top\mathbf{w}+\lambda\mathbf{1}\big)-\lambda\}.$$

By definition, $f^*(y_i\mathbf{w}) = \sup_{\boldsymbol{\delta}\in\mathbb{R}^n}\big\{y_i\mathbf{w}^\top\boldsymbol{\delta}-f(\boldsymbol{\delta})\big\}$, which equals

$$\begin{aligned}\text{maximize on } \boldsymbol{\delta},\mathbf{q}: \quad & y_i\mathbf{w}^\top\boldsymbol{\delta}-\alpha(\mathbf{q})\\ \text{subject to:} \quad & \Delta_i\mathbf{q}-\boldsymbol{\delta}=0,\\ & \mathbf{1}^\top\mathbf{q}=1\\ & \mathbf{q}\geq\mathbf{0}.\end{aligned} \tag{26}$$

Notice that (26) equals

$$\mathcal{L}(\boldsymbol{\delta},\mathbf{q},\mathbf{c},\lambda) \quad \triangleq \quad \max_{\boldsymbol{\delta};\,\mathbf{q}\geq\mathbf{0}}\ \min_{\mathbf{c},\lambda}\big\{y_i\mathbf{w}^\top\boldsymbol{\delta}-\alpha(\mathbf{q})+\mathbf{c}^\top\Delta_i\mathbf{q}-\mathbf{c}^\top\boldsymbol{\delta}+\lambda\mathbf{1}^\top\mathbf{q}-\lambda\big\}.$$

Since Problem (26) is convex and all constraints are linear, Slater's condition is satisfied and the duality gap is zero. Hence, we can exchange the order of minimization and maximization:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\delta},\mathbf{q},\mathbf{c},\lambda) &= \min_{\mathbf{c},\lambda}\max_{\boldsymbol{\delta},\mathbf{q}\geq\mathbf{0}}\big\{y_i\mathbf{w}^\top\boldsymbol{\delta}-\alpha(\mathbf{q})+\mathbf{c}^\top\Delta_i\mathbf{q}-\mathbf{c}^\top\boldsymbol{\delta}+\lambda\mathbf{1}^\top\mathbf{q}-\lambda\big\}\\ &= \min_{\mathbf{c},\lambda}\big\{\max_{\boldsymbol{\delta}}\big(y_i\mathbf{w}^\top\boldsymbol{\delta}-\mathbf{c}^\top\boldsymbol{\delta}\big)+\max_{\mathbf{q}\geq\mathbf{0}}\big(\mathbf{c}^\top\Delta_i\mathbf{q}+\lambda\mathbf{1}^\top\mathbf{q}-\alpha(\mathbf{q})\big)-\lambda\big\}\\ &= \min_{\lambda}\big\{\max_{\mathbf{q}\geq\mathbf{0}}\big(y_i\mathbf{w}^\top\Delta_i\mathbf{q}+\lambda\mathbf{1}^\top\mathbf{q}-\alpha(\mathbf{q})\big)-\lambda\big\}\\ &= \min_{\lambda}\alpha^*\big(y_i\Delta_i^\top\mathbf{w}+\lambda\mathbf{1}\big)-\lambda.\end{aligned}$$

The third equality holds because $\mathbf{c}=y_i\mathbf{w}$ is the necessary condition to make $\max_{\boldsymbol{\delta}}\big(y_i\mathbf{w}^\top\boldsymbol{\delta}-\mathbf{c}^\top\boldsymbol{\delta}\big)$ finite. ■

Example. Let $\alpha(\mathbf{q}) = \sum_{j=1}^{t} q_j \log(q_j/p_j)$, the KL divergence for discrete probability measures. By applying Theorem 17, Constraint (25) is equivalent to

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - \log\big(\sum_{j=1}^{t} p_j \exp(y_i \mathbf{w}^\top \boldsymbol{\delta}_i^j)\big) \geq 1 - \xi_i,$$

$$\iff \quad \sum_{j=1}^{t} p_j \exp\Big(y_i \mathbf{w}^\top \boldsymbol{\delta}_i^j - y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) + 1 - \xi_i\Big) \leq 1.$$

This is a geometric program, which is a well-studied class of convex problems with specialized and efficient algorithms for their solution (Boyd and Vandenberghe, 2004).

5. Kernelized Comprehensive Robust Classifier

Much of the previous development can be extended to the kernel space. The main contributions in this section are (i) a representer theorem in the case where we have discount functions in the feature space; and (ii) a sufficient condition for approximation in the case that we have discount functions in the original sample space.

We use $k(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ to represent the kernel function, and K to denote the Gram matrix with respect to $(\mathbf{x}_1, \cdots, \mathbf{x}_m)$. We assume that K is a non-zero matrix without loss of generality.

We first investigate the case where the noise exists explicitly in the feature space. Let $\phi(\cdot)$ be the mapping from the sample space $\mathbb{R}^n$ to the feature space Φ. Let $\hat{\Phi} \subseteq \Phi$ be the subspace spanned by $\{\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_m)\}$. For a vector $\mathbf{z} \in \Phi$, denote $\mathbf{z}^=$ as its projection on $\hat{\Phi}$, and $\mathbf{z}^\perp \triangleq \mathbf{z} - \mathbf{z}^=$ as its residual. The following theorem states that we can focus on $\mathbf{w} \in \hat{\Phi}$ without loss of generality.

Theorem 18 *If $f_i(\cdot)$ is such that*

$$f_i(\boldsymbol{\delta}) \geq f_i(\boldsymbol{\delta}^=), \quad \forall \boldsymbol{\delta} \in \Phi,$$

and $\mathbf{w} \in \Phi$ satisfies

$$y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \Phi, \tag{27}$$

then its projection $\mathbf{w}^=$ also satisfies (27).

Proof Before proving this theorem, we first establish the following two lemmas.

Lemma 19 *If $\mathbf{w} \in \Phi$ satisfies*

$$y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \hat{\Phi}, \tag{28}$$

then its projection $\mathbf{w}^=$ also satisfies (28).

Proof Decompose $\mathbf{w} = \mathbf{w}^= + \mathbf{w}^\perp$. By definition, $\mathbf{w}^\perp$ is orthogonal to $\hat{\Phi}$. Since $\boldsymbol{\delta}_i \in \hat{\Phi}$ and $\phi(\mathbf{x}_i) \in \hat{\Phi}$, we have

$$\langle \mathbf{w}^\perp,\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i\rangle = 0, \quad \forall \boldsymbol{\delta}_i \in \hat{\Phi},$$

which establishes the lemma. ■

Lemma 20 *If $f_i(\cdot)$ is such that*

$$f_i(\boldsymbol{\delta}) \geq f_i(\boldsymbol{\delta}^{=}), \quad \forall \boldsymbol{\delta} \in \Phi,$$

and $\mathbf{w} \in \hat{\Phi}$ satisfies

$$y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \hat{\boldsymbol{\delta}}_i \rangle + b) \geq 1 - \xi_i - f_i(\hat{\boldsymbol{\delta}}_i), \quad \forall \hat{\boldsymbol{\delta}}_i \in \hat{\Phi}, \tag{29}$$

then $\mathbf{w}$ satisfies

$$y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \Phi. \tag{30}$$

Proof We prove this lemma by deriving a contradiction. Assume that there exists $\boldsymbol{\delta}' \in \Phi$ such that Inequality (30) does not hold, i. e.,

$$y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}' \rangle + b) < 1 - \xi_i - f_i(\boldsymbol{\delta}').$$

Decompose $\boldsymbol{\delta}' = \boldsymbol{\delta}'^{=} + \boldsymbol{\delta}'^{\perp}$. Hence we have $f_i(\boldsymbol{\delta}'^{=}) \leq f_i(\boldsymbol{\delta}')$ by assumption, and $\langle \mathbf{w},\, \boldsymbol{\delta}'^{\perp} \rangle = 0$ since $\mathbf{w} \in \hat{\Phi}$. This leads to

$$\begin{aligned} y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}'^{=} \rangle + b) &= y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}' \rangle + b) \\ &< 1 - \xi_i - f_i(\boldsymbol{\delta}') \leq 1 - \xi_i - f_i(\boldsymbol{\delta}'^{=}), \end{aligned}$$

which contradicts (29) and hence we prove the lemma. ■

Now we proceed to prove Theorem 18. Since $\mathbf{w}$ satisfies (27), then it also satisfies

$$y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \hat{\Phi}.$$

Thus by Lemma 19, $\mathbf{w}^{=}$ satisfies

$$y(\langle \mathbf{w}^{=},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \hat{\Phi}.$$

By Lemma 20, this implies that $\mathbf{w}^{=}$ satisfies

$$y(\langle \mathbf{w}^{=},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \Phi,$$

which establishes the theorem. ■

The kernelized comprehensive robust classifier can be written as:

Kernelized Comprehensive Robust Classifier:

$$\begin{aligned} \min: \quad & r\big(\sum_{j=1}^{m} \alpha_j \phi(\mathbf{x}_j), b\big) + \sum_{i=1}^{m} \xi_i, \\ \text{s.t.}: \quad & y_i(\langle \sum_{j=1}^{m} \alpha_j \phi(\mathbf{x}_j),\, \phi(\mathbf{x}_i) - \sum_{j=1}^{m} c_j \phi(\mathbf{x}_j) \rangle + b) \geq \\ & \qquad 1 - \xi_i - f_i\big(\sum_{j=1}^{m} c_j \phi(\mathbf{x}_j)\big), \quad \forall \big(c_1, \cdots, c_m\big) \in \mathbb{R}^m,\ i = 1, \cdots, m, \\ & \xi_i \geq 0,\ i = 1, \cdots, m, \end{aligned} \tag{31}$$

Define $\mathbf{c} \triangleq (c_1, \cdots, c_m)$, $g_i(\mathbf{c}) \triangleq f_i(\sum_{i=1}^m c_i \phi(\mathbf{x}_i))$, and $\tilde{r}(\boldsymbol{\alpha}, b) \triangleq r\big(\sum_{j=1}^m \alpha_j \phi(\mathbf{x}_j), b\big)$. Let $\mathbf{e}_i$ denote the i^{th} basis vector. Then Problem (31) can be rewritten as

$$\begin{aligned} \min: \quad & \tilde{r}(\boldsymbol{\alpha}, b) + \sum_{i=1}^m \xi_i, \\ \text{s.t.}: \quad & y_i(\mathbf{e}_i^\top K\boldsymbol{\alpha} + b) - y_i \boldsymbol{\alpha}^\top K\mathbf{c} \geq 1 - \xi_i - g_i(\mathbf{c}), \qquad \forall \mathbf{c} \in \mathbb{R}^m,\ i = 1, \cdots, m, \\ & \xi_i \geq 0,\ i = 1, \cdots, m, \end{aligned}$$

where the constraint can be further simplified as

$$y_i(\mathbf{e}_i^\top K\boldsymbol{\alpha} + b) - g_i^*(y_i K\boldsymbol{\alpha}) \geq 1 - \xi_i,\ i = 1, \cdots, m.$$

Notice that generally $g^*(\cdot)$ depends on the exact formulation of the feature mapping $\phi(\cdot)$. However, for the following specific class of $f(\cdot)$, we can determine $g^*(\cdot)$ from K without knowing $\phi(\cdot)$.

Theorem 21 *If there exists $h_i : \mathbb{R}^+ \to \mathbb{R}^+$ such that*

$$f_i(\boldsymbol{\delta}) = h_i(\sqrt{\langle \boldsymbol{\delta}, \boldsymbol{\delta} \rangle}), \forall \boldsymbol{\delta} \in \Phi,$$

then

$$g_i^*(y_i K\alpha) = h_i^*(\|\boldsymbol{\alpha}\|_K).$$

Proof By definition,

$$\begin{aligned} g_i^*(y_i K\alpha) &= \sup_{\mathbf{c} \in \mathbb{R}^m} \Big\{ y_i \boldsymbol{\alpha}^\top K\mathbf{c} - g_i(\mathbf{c}) \Big\} \\ &= \sup_{\mathbf{c} \in \mathbb{R}^m} \left\{ y_i \boldsymbol{\alpha}^\top K\mathbf{c} - f_i\big(\sum_{j=1}^m c_j \phi(\mathbf{x}_j)\big) \right\} \\ &= \sup_{\mathbf{c} \in \mathbb{R}^m} \left\{ y_i \boldsymbol{\alpha}^\top K\mathbf{c} - h_i\left(\sqrt{\langle \sum_{j=1}^m c_j \phi(\mathbf{x}_j), \sum_{j=1}^m c_j \phi(\mathbf{x}_j) \rangle} \right) \right\} \\ &= \sup_{\mathbf{c} \in \mathbb{R}^m} \Big\{ y_i \boldsymbol{\alpha}^\top K\mathbf{c} - h_i(\sqrt{\mathbf{c}^\top K\mathbf{c}}) \Big\}. \end{aligned}$$

Notice the right-hand side can be written as

$$\begin{aligned} \sup_{\mathbf{c} \in \mathbb{R}^m} \Big\{ (y_i K^{1/2}\boldsymbol{\alpha})^\top (K^{1/2}\mathbf{c}) - h_i(\|K^{1/2}\mathbf{c}\|_2) \Big\} &= \sup_{\mathbf{c} \in \mathbb{R}^m} \Big\{ \|y_i K^{1/2}\boldsymbol{\alpha}\|_2 \|K^{1/2}\mathbf{c}\|_2 - h_i(\|K^{1/2}\mathbf{c}\|_2) \Big\} \\ &= \sup_{s \in \mathbb{R}^+} \Big\{ s\|(K^{1/2}\boldsymbol{\alpha})\|_2 - h_i(s) \Big\} = h_i^*(\|\boldsymbol{\alpha}\|_K). \end{aligned}$$

Here, the first equality holds since

$$(y_i K^{1/2}\boldsymbol{\alpha})^\top (K^{1/2}\mathbf{c}) \leq \|y_i K^{1/2}\boldsymbol{\alpha}\|_2 \|K^{1/2}\mathbf{c}\|_2$$

by Hölder's inequality. And the equality can be reached by taking $\mathbf{c}$ equal to $y_i\boldsymbol{\alpha}$ multiplied by a constant. The third equality holds because when K is non-zero, $\|K^{1/2}\mathbf{c}\|_2$ ranges over

$\mathbb{R}^+$. ■

Notice that when h_i is an increasing function, then $f_i(\boldsymbol{\delta}) \geq f_i(\boldsymbol{\delta}^=)$ is automatically satisfied $\forall \boldsymbol{\delta} \in \Phi$.

The previous results hold for the case where we have explicit discount functions in the feature space. However, in certain cases the discount functions naturally lie in the original sample space. The next theorem gives a sufficient alternative in this case.

Theorem 22 *Suppose* $h_i : \mathbb{R}^+ \to \mathbb{R}^+$ *satisfies*

$$h_i\big(\sqrt{k(\mathbf{x}_i, \mathbf{x}_i) + k(\mathbf{x}_i - \boldsymbol{\delta}, \mathbf{x}_i - \boldsymbol{\delta}) - 2k(\mathbf{x}_i, \mathbf{x}_i - \boldsymbol{\delta})}\big) \leq f_i(\boldsymbol{\delta}), \quad \forall \boldsymbol{\delta} \in \mathbb{R}^n. \tag{32}$$

Then

$$y_i\big(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_{\boldsymbol{\phi}} \rangle + b\big) \geq 1 - \xi_i - h_i(\sqrt{\langle \boldsymbol{\delta}_{\boldsymbol{\phi}}, \boldsymbol{\delta}_{\boldsymbol{\phi}} \rangle}), \quad \forall \boldsymbol{\delta}_{\boldsymbol{\phi}} \in \Phi, \tag{33}$$

implies

$$y_i\big(\langle \mathbf{w}, \phi(\mathbf{x}_i - \boldsymbol{\delta}) \rangle + b\big) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}), \quad \forall \boldsymbol{\delta} \in \mathbb{R}^n. \tag{34}$$

Proof Notice that (33) implies that

$$\begin{aligned} y_i\Big(\big\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \big[\phi(\mathbf{x}_i) - \phi(\mathbf{x}_i - \boldsymbol{\delta})\big] \big\rangle + b\Big) \\ \geq 1 - \xi_i - h_i\big(\sqrt{\big\langle \phi(\mathbf{x}_i) - \phi(\mathbf{x}_i - \boldsymbol{\delta}),\, \phi(\mathbf{x}_i) - \phi(\mathbf{x}_i - \boldsymbol{\delta}) \big\rangle}\big), \end{aligned}$$

$\forall \boldsymbol{\delta} \in \mathbb{R}^n$. The right-hand side is equal to

$$1 - \xi_i - h_i\big(\sqrt{k(\mathbf{x}_i,\, \mathbf{x}_i) + k(\mathbf{x}_i - \boldsymbol{\delta}, \mathbf{x}_i - \boldsymbol{\delta}) - 2k(\mathbf{x}_i, \mathbf{x}_i - \boldsymbol{\delta})}\big) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}).$$

Since this holds for all $\boldsymbol{\delta} \in \mathbb{R}^n$, (34) holds for $(\mathbf{w},\, b)$. ■

Notice the condition in Theorem 22 only involves the kernel function $k(\cdot, \cdot)$ and is independent of the explicit feature mapping. Hence this theorem applies for abstract mappings, and specifically mappings into infinite-dimension spaces.

Theorem 23 *Equip the sample space with a metric* $d(\cdot, \cdot)$*, and suppose there exist* $\hat{k}_i : \mathbb{R}^+ \to \mathbb{R}$*, and* $\hat{f}_i : \mathbb{R}^+ \to \mathbb{R} \bigcup \{+\infty\}$ *such that,*

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \hat{k}(d(\mathbf{x}, \mathbf{x}')), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^n; \\ f_i(\boldsymbol{\delta}) &= \hat{f}_i(d(\mathbf{x}_i, \mathbf{x}_i - \boldsymbol{\delta}_i)), \forall \boldsymbol{\delta} \in \mathbb{R}^n. \end{aligned} \tag{35}$$

Then $h_i : \mathbb{R}^+ \to \mathbb{R}^+ \bigcup \{+\infty\}$ *defined as*

$$h_i(x) = \inf_{y \mid \exists \mathbf{z} \in \mathbb{R}^n : y = d(\mathbf{x}_i, \mathbf{z}),\, \hat{k}(y) = \hat{k}(0) - x^2/2} \hat{f}_i(y) \tag{36}$$

satisfies Equation (32), and for any $h'(\cdot)$ *that satisfies Equation (32),* $h'(x) \leq h(x), \forall x \geq 0$ *holds. Here, we take* $\inf_{y \in \emptyset} \hat{f}_i(y)$ *to be* $+\infty$*.*

Proof Rewrite Inequality (32) as

$$h_i\big(\sqrt{\hat{k}(d(\mathbf{x}_i,\mathbf{x}_i))+\hat{k}(d(\mathbf{x}_i-\boldsymbol{\delta},\mathbf{x}_i-\boldsymbol{\delta}))-2k(d(\mathbf{x}_i,\mathbf{x}_i-\boldsymbol{\delta}))}\big)\leq \hat{f}_i(d(\mathbf{x}_i,\mathbf{x}_i-\boldsymbol{\delta})),\quad \forall\boldsymbol{\delta}\in\mathbb{R}^n,$$

which is equivalent to

$$h_i(x)\leq \hat{f}_i(y);\quad \forall x,y,\mathbf{z}: x=\sqrt{2\hat{k}(0)-y};\ y=d(\mathbf{x}_i,\mathbf{z}).$$

Observe that the function defined by (36) is the maximal function that satisfies this inequality, thus proving the theorem. ■

Remark 24 In most cases, $\hat{f}_i$ is increasing and piecewise continuous, $d(\cdot,\cdot)$ satisfies that for any $y\geq 0$, there exists $\mathbf{z}\in\mathbb{R}^n$ such that $d(\mathbf{x}_i,\mathbf{z})=y$. Equation (36) can be simplified as

$$h_i(x)=\begin{cases}+\infty & \{y|\hat{k}(y)=\hat{k}(0)-x^2/2\}\text{ is empty}\\ \hat{f}_i\Big(\hat{k}^{-1}\big(\hat{k}(0)-x^2/2\big)\Big) & \min\{y|\hat{k}(y)=\hat{k}(0)-x^2/2\}\text{ exists}\\ \hat{f}_i\Big(\hat{k}^{-1}\big(\hat{k}(0)-x^2/2\big)^+\Big) & \text{otherwise.}\end{cases}$$

Here, $\hat{k}^{-1}(x)\triangleq\inf\{y|\hat{k}(y)=x\}$, and $\hat{f}_i(c^+)$ stands for the right limit at c of $f_i(\cdot)$.

Consider the Gaussian Kernel $k(\mathbf{x},\mathbf{x}')=\exp(-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2)$ as an example. We have $d(\mathbf{x},\mathbf{x}')=\|\mathbf{x}-\mathbf{x}'\|$ and $\hat{k}(x)=\exp(-x^2/2\sigma^2)$. Hence $\hat{k}^{-1}(y)=\sqrt{-2\sigma^2\ln y}$ yields

$$h_i(x)=\begin{cases}\hat{f}_i\Big(\sqrt{-2\sigma^2\ln(1-x^2/2)}\Big) & x\leq\sqrt{2}\\ +\infty & \text{otherwise.}\end{cases}$$

Taking $\hat{f}_i(x)=I_c$, the corresponding $h_i(x)=I_{2-2\exp(-c^2/2\sigma^2)}$. Taking $\hat{f}_i(x)=cx^2$, the corresponding $h_i(x)$ is

$$h_i(x)=\begin{cases}-2c\sigma^2\ln(1-x^2/2) & x\leq\sqrt{2}\\ +\infty & \text{otherwise.}\end{cases}$$

6. Numerical Simulations

In this section, we use empirical experiments to gain further insight into the performance of the comprehensive robust classifier. To this end, we compare the performance of three classification algorithms: the standard SVM, the standard robust SVM with ellipsoidal uncertainty set, and comprehensive robust SVM with ellipsoidal uncertainty set with linear discount function from the center of the ellipse to its boundary (see below). The simulation results show that a comprehensive robust classifier with the discount function appropriately tuned has a performance superior to both the robust classifier and the standard SVM. The empirical results show that this soft formulation of robustness builds in protection to noise, without being overly conservative.

We use the non-kernelized version for both the robust classification and the comprehensive robust classification. We use a linear discount function for the comprehensive robust

classifier. That is, noise is bounded in the same ellipsoidal set as for the robust SVM, $\{\boldsymbol{\delta} | \|\boldsymbol{\delta}\|_{\Sigma^{-1}} \leq 1\}$, and the discount function is

$$f_i(\boldsymbol{\delta}) = \begin{cases} \alpha \|\boldsymbol{\delta}\|_{\Sigma^{-1}} & \|\boldsymbol{\delta}\|_{\Sigma^{-1}} \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

The parameter α controls the disturbance discount. As α tends to zero, there is no discount inside the uncertainty set, and we recover the robust classifier. As α tends to $+\infty$, the discount increases until effectively the constraint is only imposed at the center of the ellipse, hence recovering the standard SVM classifier.

We use SeduMi 1.1R3 (Sturm, 1999) to solve the resulting convex programs. We first compare the performance of the three algorithms on the Wisconsin-Breast-Cancer data set from the UCI repository (Asuncion and Newman). In each iteration, we randomly pick 50% of the samples as training samples and the rest as testing samples. Each sample is corrupted by i.i.d. noise, which is uniformly distributed in an ellipsoid $\{\boldsymbol{\delta} | \|\boldsymbol{\delta}\|_{\Sigma^{-1}} \leq 1\}$. Here, the matrix Σ is diagonal. For the first 40% of features, $\Sigma_{ii} = 16$, and for the remaining features, $\Sigma_{ii} = 1$. This ellipsoidal uncertainty set captures the setup where noise is skewed toward part of the features. This is more common in practice than the case where the noise is equal for all features. We repeat 30 such iterations to get the average empirical error of the three different algorithms. Figure 3 shows that for appropriately chosen discount parameter α, the comprehensive robust classifier outperforms both the robust and standard SVM classifiers. As anticipated, when α is small, comprehensive robust classification has a testing error rate comparable to robust classification. For large α, the classifier's performance is similar to that of the standard SVM. This figure essentially shows that protection against noise is beneficial as long as it does not become overly conservative. Comprehensive robust classification is of interest because it provides a more flexible approach to handle the noise.

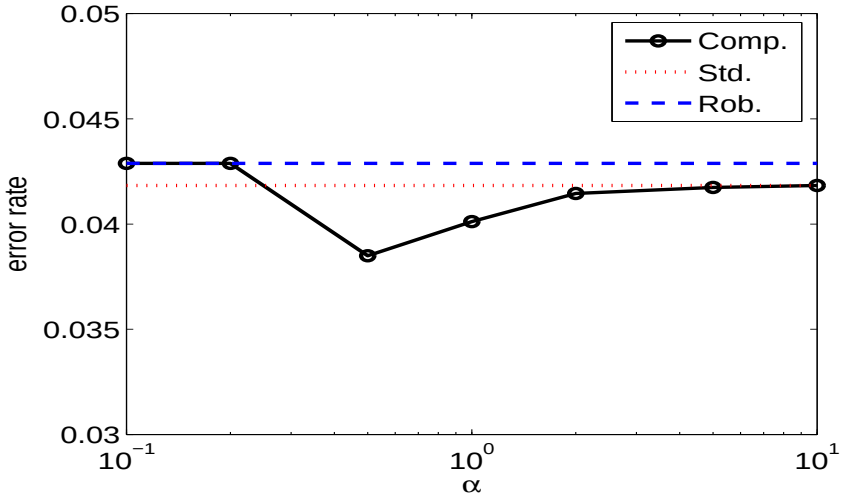


Figure 3: Simulation results for WBC Data.

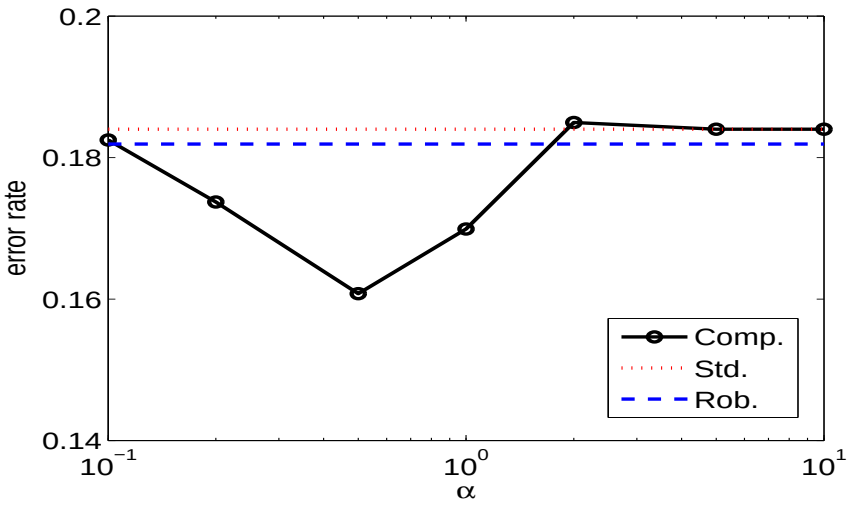


Figure 4: Simulation results for Ionosphere Data.

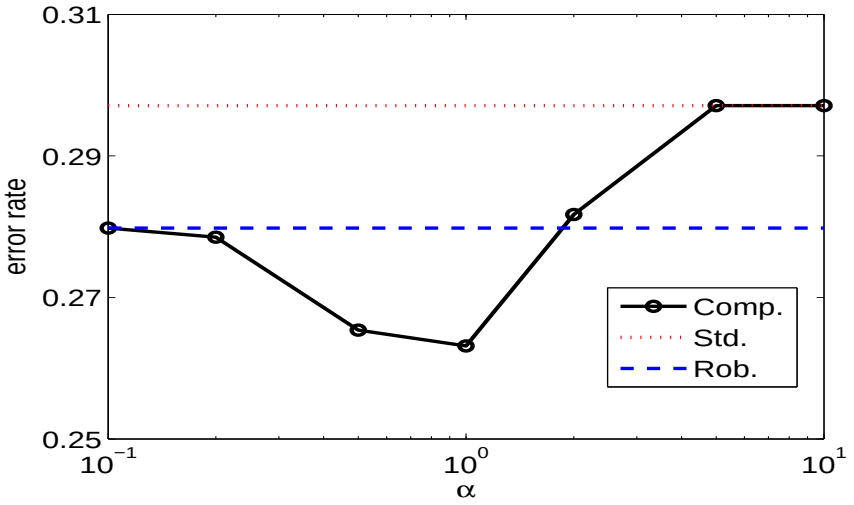


Figure 5: Simulation results for Sonar Data.

We run similar simulations on Ionosphere and Sonar data sets from the UCI repository Asuncion and Newman. To fit the variability of the data, we scale the uncertainty set: for 40% of the features, Σ_{ii} equals 0.3 for Ionosphere and 0.01 for Sonar; for the remaining features, Σ_{ii} equals 0.0003 for Ionosphere and 0.00001 for Sonar. Figure 4 and Figure 5 show the respective simulation results. Similarly to the WBC data set, comprehensive robust classification achieves its optimal performance for mid-range α, and is superior to both the standard SVM and the robust SVM.

7. Concluding Remarks

This work investigates the relationship between robust classification and its extensions, and regularized SVM classification, and seeks to develop robust classifiers with controlled conservatism. In particular, we show that the standard norm-regularized SVM classifier is in fact the solution to a robust classification setup, and thus known results about regularized classifiers extend to robust classifiers. To the best of our knowledge, this is the first explicit such link between regularization and robustness in pattern classification. This link suggests that norm-based regularization essentially builds in a robustness to sample noise whose probability level sets are symmetric, and moreover have the structure of the unit ball with respect to the dual of the regularizing norm. It would be interesting to understand the performance gains possible when the noise does not have such characteristics, and the robust setup is used in place of regularization with appropriately defined uncertainty set.

We further expand on this connection by showing that any arbitrary convex constraint regularization is equivalent to the classifier obtained through a formulation using a softer version of robustness known as comprehensive robustness. This allows the connection to convex risk measures, from which we develop risk-constrained classifiers.

At a high level, our contribution is the introduction of a more geometric notion of hedging and controlling complexity (robust and comprehensive robust classifiers integrally depend on the uncertainty set and structure of the discount function) and the link to probabilistic notions of hedging, including chance constraints and convex risk constraints. We believe that in the realm of applications, particularly when distribution-free PAC-style bounds are pessimistic, the design flexibility of such a framework will yield superior performance. A central issue on the application front is to understand how to effectively use the additional degrees of freedom and flexibility since now we are designing uncertainty sets and discount functions, rather than simply choosing regularization parameters that multiply a norm.

Appendix A.

In this appendix we prove **Theorem 14**. Before proving Theorem 14, we establish the following two lemmas. Lemma 25 is adapted from Fóllmer and Schied (2002), and the readers can find the proof there.

Lemma 25 *Let $\mathcal{X}$ be the set of random variables for $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{P}$ be the set of probability measures absolutely continuous with respect to $\mathbb{P}$, and $\rho : \mathcal{X} \to \mathbb{R}$ be a convex risk measure satisfying $X_n \downarrow X \Rightarrow \rho(X_n) \to \rho(X)$, then there exists a convex function $\alpha : \mathcal{P} \to (-\infty, +\infty]$ such that*

$$\rho(X) = \sup_{Q \in \mathcal{P}} \big(\mathbb{E}_Q(X) - \alpha(Q) \big) \quad \forall X \in \mathcal{X}. \tag{37}$$

Furthermore, $\alpha^0(Q) \triangleq \sup_{X' \in \mathcal{X}} \big(\mathbb{E}_Q(X') - \rho(X') \big)$ satisfies (37), and it is minimal in the sense that $\alpha^0(Q) \leq \alpha(Q)$ for all $Q \in \mathcal{P}$, if $\alpha(\cdot)$ also satisfies (37).

We call $\alpha^0(\cdot)$ the minimal representation of a convex risk measure.

Lemma 26 *For a normalized convex risk measure $\rho(\cdot)$, its minimal representation satisfies:*

$$0 = \alpha^0(\mathbb{P}) \leq \alpha^0(Q), \forall Q \ll \mathbb{P}.$$

Proof First, since $\mathbb{E}_Q(0) \equiv 0$, we have

$$\rho(0) = 0 \to \inf_{Q \in \mathcal{P}} \alpha^0(Q) = 0. \tag{38}$$

Next, by definition $\alpha^0(\mathbb{P}) = \sup_{X \in \mathcal{X}} \big(\mathbb{E}_{\mathbb{P}}(X) - \rho(X) \big)$, and $\mathbb{E}_{\mathbb{P}}(X) \leq \rho(X)$ by assumption. Hence taking the supremum leads to $\alpha_0(\mathbb{P}) \leq 0$. Combining this with Equation (38) establishes the lemma. ■

Now we proceed to prove Theorem 14.

Proof

1. By Lemma 26, $f_i(\boldsymbol{\delta}_i) \geq 0$ since $\alpha^0(Q) \geq 0$, $\forall Q \in \mathcal{P}$. In addition, $\mathbb{E}_{\mathbb{P}}(\boldsymbol{\delta}_i) = \mathbf{0}$ and $\alpha^0(\mathbb{P}) = 0$ together imply $f_i(\mathbf{0}) = 0$. Hence $f_i(\cdot)$ satisfies (14).

   Inequality (25) can be rewritten as

$$\begin{aligned}
& \xi_i + y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq \sup_{Q \in \mathcal{P}} \big(\mathbb{E}_Q(y_i \mathbf{w}^\top \boldsymbol{\delta}_i^r) - \alpha(Q) \big) \\
\iff \quad & \xi_i + y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq \sup_{\boldsymbol{\delta}_i \in \mathbb{R}^n} \sup_{Q \in \mathcal{P} | \mathbb{E}_Q(\boldsymbol{\delta}_i^r) = \boldsymbol{\delta}_i} \big(y_i \mathbf{w}^\top \boldsymbol{\delta}_i - \alpha(Q) \big) \\
\iff \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - \inf\{\alpha(Q) | \mathbb{E}_Q(\boldsymbol{\delta}_i^r) = \boldsymbol{\delta}_i\}, \ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n, \\
\iff \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n,
\end{aligned}$$

   which proves the first part.

2. First we show $\rho_i(\cdot)$ is a convex risk measure. Notice $f_i(\mathbf{0})$ is finite, hence, $\rho_i(X) > -\infty$. Observe that $\rho_i(\cdot)$ satisfies Translation Invariance. To prove Monotonicity, suppose $X \leq Y$ and $Y - s \in \mathcal{A}_i$ for some $s \in \mathbb{R}$, then $X - s \in \mathcal{A}_i$, hence $\inf\{m | X - m \in \mathcal{A}_i\} \leq s$, which implies $\rho_i(X) \leq \rho_i(Y)$. To prove Convexity, suppose $X - m$ and $Y - n$ belong to $\mathcal{A}_i$ for $m, n \in \mathbb{R}$. Given $\lambda \in [0,1]$, we have $\lambda(X(\omega) - m) + (1-\lambda)(Y(\omega) - n) \leq f_i\big(\boldsymbol{\delta}_i^r(\omega)\big)$ and hence $\big(\lambda X + (1-\lambda)Y\big) - (\lambda m + (1-\lambda)n) \in \mathcal{A}_i$ which implies $\rho_i(\lambda X + (1-\lambda)Y) \leq \lambda m + (1-\lambda)n$, hence the convexity holds. Therefore $\rho_i(\cdot)$ is a convex risk measure.

   Inequality (25) can be rewritten as

$$\begin{aligned}
&\inf\{m \in \mathbb{R} | y_i \mathbf{w}^\top \boldsymbol{\delta}_i^r - m \in \mathcal{A}_i\} \leq \xi_i + y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \\
\iff \quad & y_i \mathbf{w}^\top \boldsymbol{\delta}_i^r - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + 1 - \epsilon \in \mathcal{A}_i, \quad \forall \epsilon > 0 \\
\iff \quad & y_i \mathbf{w}^\top \boldsymbol{\delta}_i^r(\omega) - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + 1 - \epsilon \leq f_i(\boldsymbol{\delta}_i^r(\omega)), \qquad \forall \omega \in \Omega, \forall \epsilon > 0 \\
\iff \quad & y_i \mathbf{w}^\top \boldsymbol{\delta}_i - \xi_i - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + 1 - \epsilon \leq f_i(\boldsymbol{\delta}_i), \qquad \forall \boldsymbol{\delta}_i \in \mathbb{R}^n.
\end{aligned}$$

   The last equivalence holds from the assumption that $\boldsymbol{\delta}_i^r$ has support $\mathbb{R}^n$.

■

For the first part of Theorem 14, the assumption that $\rho_i(\cdot)$ is normalized can be relaxed to $\rho_i(0) = 0$ and $\inf\{\alpha_i^0(Q) | \mathbb{E}_Q(\boldsymbol{\delta}_i^r) = \mathbf{0}\} = 0$.

References

A. Asuncion and D.J. Newman. UCI machine learning repository.

P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, November 2002.

P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexity. *The Annals of Statictics*, 33(4):1497–1537, 2005.

A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, August 1999.

A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust solutions of uncertain linear programs. *Math. Programming*, 99:351–376, 2003.

A. Ben-Tal, D. Bertsimas, and D. Brown. A flexible approach to robust optimization via convex risk measures. Submiteed at Sep. 2006, September 2006a.

A. Ben-Tal, S. Boyd, and A. Nemirovski. Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems. *Mathematical Programming, Series B*, 107, 2006b.

K. Bennett and O. Mangasrian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1):23–34, 1992.

D. Bertsimas and D. B. Brown. Robust linear optimization and coherent rish measures. Technical Report LIDS #2659, Massachusetts Institute of Technology, March 2005.

D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, January 2004.

C. Bhattacharyya, L. Grate, M. Jordan, L. El-Ghaoui, and I. Mian. Robust sparse hyperplane classifiers: Application to uncertain molecular profiling data. *Journal of Computational Biology*, 11(6):1073–1089, 2004a.

C. Bhattacharyya, K. Pannagadatta, and A. Smola. A second order cone programming formulation for classifying missing data. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems (NIPS17)*, Cambridge, MA, 2004b. MIT Press.

P. Boser, I. Guyon, and V. Vapnik. A traing algorithm for optimal margni classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, New York, NY, 1992.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. In A. Smola, P. Bartlett, B. Schólkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 171–203, Cambridge, MA, 2000. MIT Press.

H. Fóllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6:429–447, 2002.

L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.

A. Globerson and S. Roweis. Nightmare at test time: robust learning by feature deletion. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, 2006. ISBN 1.

M. Grótschel, L. Lovasz, and A. Schrijver. *The Ellipsoid Method and Combinatorial Optimization*. Springer, Heidelberg, 1988.

M. Kearns, Y. Mansour, A. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1C50, 2002.

G. Lanckriet, L. El-Ghaoui, C. Bhattacharyya, and M. Jordan. A robust minimax approach to classfication. *Journal of Machine Learning Research*, 3:555–582, December 2002.

A. Marshall and I. Olkin. Multivariate chebyshev inequalities. *Annuals of Mathematical Statistics*, 31(4):1001, 1960.

R.T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, N. J., 1970.

P. Shivaswamy, C. Bhattacharyya, and A. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, July 2006.

A. Smola, B. Schólkopf, and K. Múllar. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Info. Theory*, 51(1):128–142, 2005.

J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999. URL `citeseer.ist.psu.edu/sturm99using.html`. Special issue on Interior Point Methods (CD supplement with software).

T. Trafalis and R. Gilbert. Robust support vector machines for classficiation and computational issues. *Optimization Methods and Software*, 22(1):187–198, February 2007.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recignition*. Nauka, Moscow, 1974.

V. Vapnik and A. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimizatin method. *Pattern Recognition and Image Analysis*, 1(3): 260–284, 1991.

V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:744–780, 1963.