Robustness, Risk, and Regularization in Support Vector Machines

Xu Huan,[*] **and Shie Mannor,**[*] **and Constantine Caramanis**[†]

Abstract

We consider two new formulations for classification problems in the spirit of support vector machines based on robust optimization. Our new 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 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.

1 Introduction

Support Vector Machines (SVMs for short), originated in [1] and can be traced back as early as [2] and [3]. 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, as suggested by [4] and [5]. 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 [6]. 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 (cf [7, 8, 9, 10, 11] and references therein).

In this paper we follow a different approach, first proposed in [12]. 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 [13, 14, 15] and references therein), so-called comprehensive robust optimization [16], and risk theory [17, 18], to derive new robust SVM classifiers. The use of robust optimization in classification is not new; see, for example, [19, 12, 20]. 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 (see, e.g., [13]). 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 the norm-

[*] Department of Electrical and Computer Engineering, McGill University, xuhuan@cim.mcgill.ca, shie@ece.mcgill.ca

[†] Department of Electrical and Computer Engineering, The University of Texas at Austin, cmcaram@ece.utexas.edu

regularized SVM classifier is *equivalent* to a robust SVM classifier.

- 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.

Structure of the Paper: This paper is organized as follows. In Section 2 the equivalence between the robust classification and the regularization process is shown. 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 other research, limits 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 suggesting 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 number of training samples $\{\mathbf{x}_i, y_i\}_{i=1}^m \subseteq \mathbb{R}^n \times \{-1,+1\}$. A linear classifier is specified by the function $h^{\mathbf{w},b}(\mathbf{x}) = \mathrm{sgn}(\langle\mathbf{w},\,\mathbf{x}\rangle + b)$. For the standard regularized classifier, the parameters $(\mathbf{w}, b)$ are obtained by solving the following convex optimization problem:

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

where $r(\mathbf{w}, b)$ is a regularization term. The standard robust optimization techniques robustify at a constraint-wise level, allowing the disturbances $\tilde{\delta} = (\delta_1,\ldots,\delta_m)$ to lie in some uncertainty set $\mathcal{N}$:

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

It is well-known (e.g., [21]) 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 δ_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 is to obtain a robust formulation where the disturbances $\{\delta_i\}$ 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:

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

and we thus obtain:

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

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

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

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*

(I) $\mathbf{0} \in \mathcal{N}_0$;

(II) $\sup_{\boldsymbol{\delta}\in\mathcal{N}_0}\big[\mathbf{w}^\top\boldsymbol{\delta}\big] = \sup_{\boldsymbol{\delta}'\in\mathcal{N}_0}\big[-\mathbf{w}^\top\boldsymbol{\delta}'\big] < \infty,\ \forall \mathbf{w}\in\mathbb{R}^n.$

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*

(I) $\{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)|\boldsymbol{\delta}_t\in\mathcal{N}_0;\ \boldsymbol{\delta}_{i\neq t}=\mathbf{0}\}\subseteq\mathcal{N},\ \forall\, t;$

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

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

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

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, $\mathcal{N}_0$ is an atomic uncertainty set and $\mathcal{N}$ is a concave correlated uncertainty set of $\mathcal{N}_0$, then the following min-max problem*

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

is equivalent to

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

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

We defer proof of the theorem to the online appendix, [**?**]. 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 corresponding atomic set for $\mathcal{N}_{\text{box}}$ is $m\mathcal{N}_0$. 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.

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

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

$$\min:\max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\in\mathcal{T}_k}\sum_{i=1}^m\max\big[1-y_i\big(\langle\mathbf{w},\ \mathbf{x}_i-\boldsymbol{\delta}_i\rangle+b\big),0\big] \tag{6}$$

$$\min:c\|\mathbf{w}\|^*+\sum_{i=1}^m\max\big[1-y_i\big(\langle\mathbf{w},\ \mathbf{x}_i\rangle+b\big),0\big]. \tag{7}$$

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 explains the widely known fact that 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 [22] and Section 6). 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, i.e., by examining the variation of the data and solving the corresponding robust classification problem. For example, one way to regularize is 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 assumption, it can be used to approximate an upper bound for a chance-constrained classifier. Suppose the disturbance $(\boldsymbol{\delta}_1^r,\cdots\boldsymbol{\delta}_m^r)$ 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 $\eta\in[0,\ 1]$,

$$\begin{aligned}\min:\ & l\\ \text{s.t.}:\ & \mu\Big\{\textstyle\sum_{i=1}^m\max\big[1-y_i(\langle\mathbf{w},\ \mathbf{x}_i-\boldsymbol{\delta}_i^r\rangle+b),0\big]\leq l\Big\}\\ &\geq 1-\eta.\end{aligned} \tag{8}$$

The formulations in [19, 20, 23] assume uncorrelated noise and require all constraints to be satisfied with high probability *simultaneously*. They find a vector $\{\xi_1,\cdots,\xi_m\}$ 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 (8) is generally non-tractable. However, we can approximate it as follows. Let

$$c^* \triangleq \inf\{\alpha|\mu(\sum_i \|\boldsymbol{\delta}_i\| \leq \alpha) \geq 1-\eta\}.$$ [1]

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

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

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

Thus (8) is upper bounded by (7) 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 section, 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 [24]. 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*

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

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

Proof: To prove this theorem, we need to establish the following lemma. 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.

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 for that for these paired samples, the total testing error is upper bounded by

$$\max_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_n) \in \mathcal{T}_N} \sum_{i=1}^n \max\big[1 - y_i\big(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b\big), 0\big]$$

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

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

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

Therefore, the average testing error is upper bounded by

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

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

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

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. 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 [16].

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 $\xi_i(\boldsymbol{\delta}_i) \triangleq \max\big[1-y_i(\langle \mathbf{w},\, \mathbf{x}_i-\boldsymbol{\delta}_i\rangle + b), 0\big]$, the robust classifier (2) can be rewritten as:

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

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 $h_i(\cdot,\cdot):\mathbb{R}\times\mathbb{R}^n\rightarrow\mathbb{R}$ satisfy $0\leq h_i(\alpha,\beta)\leq h_i(\alpha,\mathbf{0})=\alpha$. 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:

$$\min_{\mathbf{w},b} \sup_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\in\mathcal{N}} \Big\{r(\mathbf{w},b)+\sum_{i=1}^m h_i\big(\xi_i(\boldsymbol{\delta}_i),\boldsymbol{\delta}_i\big)\Big\}, \qquad (9)$$

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 3.4 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 3.2.

We formulate the comprehensive robust classification with additive discount function in Section 3.1 and establish an equivalence relationship between comprehensive robust classifications 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 3.2 we investigate the tractability of comprehensive robust classification. In Section 3.3 we discuss a special class of discounts, namely norm discounts, and derive probability bounds for such discounts. Finally, in Section 3.4 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.[2] Substituting $h_i(\alpha,\boldsymbol{\beta})\triangleq\max(0,\alpha-f_i(\boldsymbol{\beta}))$ into Equation (9) 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:

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

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

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} f_i(\boldsymbol{\beta})=f_i(\mathbf{0})=0. \qquad (10)$$

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 $\mathbf{x}_i$ (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 $f_i(\boldsymbol{\delta}_i)=\alpha\|\boldsymbol{\delta}_i\|$.

Given a function $f(\cdot)$, let f^* denote its Legendre-Fenchel transform or conjugate function, given by $f^*(s)=\sup_x\{\langle s,y\rangle-f(x)\}$ [25]. Then we have the following, that shows that if f is a disturbance discount that satisfies (10), 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 (10), then so does $f^*(\cdot)$.*

(ii) If $g(\cdot)$ is closed and convex, and $g^*(\cdot)$ satisfies (10), then so does $g(\cdot)$.

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

Theorem 7 *The Comprehensive Robust Classifier (10) is equivalent to the following convex program:*

$$\begin{aligned} \min :& r(\mathbf{w}, b) + \sum_{i=1}^{m} \xi_i, \\ \text{s.t. :}& y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - f_i^*(y_i\mathbf{w}) \geq 1 - \xi_i,\ \ i = 1, \cdots, m, \\ & \xi_i \geq 0,\ \ i = 1, \cdots, m. \end{aligned} \tag{11}$$

Proof: Simple algebra yields

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

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,

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

which is the indicator function of the dual-norm ball with radius α. Thus (11) is equivalent to

$$\begin{array}{rl} \min : & \sum_{i=1}^m \xi_i, \\ \text{s.t. :} & y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i,\ \ i = 1, \cdots, m, \\ & \|\mathbf{w}\|^* \leq \alpha, \\ & \xi_i \geq 0,\ \ i = 1, \cdots, m. \end{array} \tag{12}$$

We notice that Problem (12) 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 comprehensive robust classifier (e.g., the probabilistic bound in Subsection 3.3) can be easily applied to robust SVMs and standard SVMs.

Tractability

We now give a sufficient condition on the discount for the comprehensive robust classification problem (11) 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*

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

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 (11) can be solved in polynomial time.

We defer the proof of this theorem to the online appendix, **[?]**.

3.3 Norm Discount

In this subsection, 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 $t_i^*(y) = \sup_{x \geq 0}\big[xy - t(x)\big]$. 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*

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - t_i(\|\boldsymbol{\delta}_i\|_{\Sigma^{-1}}),\ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n, \tag{13}$$

is equivalent to

$$\inf_{\boldsymbol{\delta}_i^r \sim (0,\Sigma)} Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) \geq 1 - \frac{1}{\big(t_i^{-1}(s)\big)^2 + 1},\ \forall s \geq 0. \tag{14}$$

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: In [19], the authors 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 (14) is equivalent to

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

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

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

Since $t_i(\cdot)$ is nondecreasing, this is equivalent to (13). ■

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,

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

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

Remark 1 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 ([26, 27, 28]) states that

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

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

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

Hence,

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

showing that the bound is tight.

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

Theorem 11 *If $\boldsymbol{\delta}_i^r \sim \mathcal{N}(0,\Sigma)$, then the constraint*

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - t_i(\|\boldsymbol{\delta}_i\|_{\Sigma^{-1}}),\ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n, \tag{15}$$

is equivalent to

$$Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) \geq \Phi\big(t^{-1}(s)\big), \tag{16}$$
$$\forall s \geq 0. \tag{17}$$

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:

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

Notice that (15) is equivalent to

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

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

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

Since $t_i(\cdot)$ is nondecreasing, this is equivalent to (15). ■

3.4 Multiplicative Discount

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

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

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

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

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

$$\begin{aligned}
& \textbf{Comprehensive Robust Classifier (Multiplicative):} \\
\min: \quad & r(\mathbf{w},b) + \textstyle\sum_{i=1}^m \xi_i, \\
\text{s.t.}: \quad & \xi_i \geq c_i(\boldsymbol{\delta})\big[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b)\big], \\
& \qquad \forall \boldsymbol{\delta}_i \in \mathbb{R}^n,\ i = 1,\cdots,m, \\
& \xi_i \geq 0,\ i = 1,\cdots,m.
\end{aligned} \tag{18}$$

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 (18) can be rewritten as:

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

Notice that 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 12 *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 subgradient.*

Then, Problem (18) can be solved in polynomial time.

We defer the proof to the online appendix, [**?**].

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 [16][29].

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

Risk-Measure Constrained Classifier

$$\begin{aligned}
\min :& r(\mathbf{w}, b) + \sum_{i=1}^{m} \xi_i, \\
\text{s.t. :}& \rho_i(\xi_i) \geq \rho_i(1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i^r \rangle + b)), \ \ i = 1, \cdots, m, \\
& \xi_i \geq 0, \ \ i = 1, \cdots, m.
\end{aligned} \tag{19}$$

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

$$\xi_i + y_i(\langle \mathbf{w},\, \mathbf{x}_i \rangle + b) - 1 \geq \rho_i(y_i \mathbf{w}^\top \boldsymbol{\delta}_i^r). \tag{20}$$

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 online appendix, online appendix, [**?**]..

Theorem 13 *(1) A Risk-Measure Constrained Classifier with normalized convex risk measures $\rho_i(\cdot)$ is equivalent to 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 [17, 18] 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 14 *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

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

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 [17].

Theorem 15 *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

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

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 3.3, we notice that the new regularizer is the result of a quadratic discount function, instead of the indicator discount function used by the 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 16 *For $\boldsymbol{\delta}_i^r$ with a finite support, the risk-measure constrained classifier is equivalent to*

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

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 (20) 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

$$\begin{aligned}
\text{Maximize on } \boldsymbol{\delta}, \mathbf{q}: \quad & y_i \mathbf{w}^\top \boldsymbol{\delta} - \alpha(\mathbf{q})\\
\text{subject to:} \quad & \Delta_i \mathbf{q} - \boldsymbol{\delta} = 0,\\
& \mathbf{1}^\top \mathbf{q} = 1\\
& \mathbf{q} \geq \mathbf{0}.
\end{aligned} \tag{21}$$

Notice that (21) equals

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

Since Problem (21) 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:

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

Here, 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 16, Constraint (20) 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, a class of convex problems and is known to be solved efficiently [27].

5 Kernelized Comprehensive Robust Classifier

Much of the previous development can be extended to the kernel space. We defer most proofs to the appendix, but give the statements of the main theorems here. 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 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 17 *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

$$y(\langle\mathbf{w},\,\phi(\mathbf{x}_i)-\boldsymbol{\delta}_i\rangle+b)\geq 1-\xi_i-f_i(\boldsymbol{\delta}_i),\quad\forall\boldsymbol{\delta}_i\in\Phi,\ (22)$$

then its projection $\mathbf{w}^{=}$ also satisfies (22).

The kernelized comprehensive robust classifier can be written as:

Kernelized Comprehensive Robust Classifier:

$$\begin{aligned}\min:\quad & r\big(\textstyle\sum_{j=1}^m\alpha_j\phi(\mathbf{x}_j),b\big)+\sum_{i=1}^m\xi_i,\\ \text{s.t.}:\quad & y_i(\langle\textstyle\sum_{j=1}^m\alpha_j\phi(\mathbf{x}_j),\,\phi(\mathbf{x}_i)-\sum_{j=1}^m c_j\phi(\mathbf{x}_j)\rangle+b)\geq\\ & \qquad 1-\xi_i-f_i\big(\textstyle\sum_{j=1}^m c_j\phi(\mathbf{x}_j)\big),\\ & \qquad\forall\big(c_1,\cdots,c_m\big)\in\mathbb{R}^m,\ i=1,\cdots,m,\\ & \xi_i\geq 0,\ i=1,\cdots,m,\end{aligned}\tag{23}$$

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 (23) can be rewritten as

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

where the constraint can be further simplified as

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

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 18 *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

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

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

$$\begin{aligned}&h_i\big(\sqrt{k(\mathbf{x}_i,\mathbf{x}_i)+k(\mathbf{x}_i-\boldsymbol{\delta},\mathbf{x}_i-\boldsymbol{\delta})-2k(\mathbf{x}_i,\mathbf{x}_i-\boldsymbol{\delta})}\big)\\ &\qquad\leq f_i(\boldsymbol{\delta}),\quad\forall\boldsymbol{\delta}\in\mathbb{R}^n.\end{aligned}\tag{24}$$

Then

$$\begin{aligned}&y_i\big(\langle\mathbf{w},\,\phi(\mathbf{x}_i)-\boldsymbol{\delta_\phi}\rangle+b\big)\geq 1-\xi_i-h_i(\sqrt{\langle\boldsymbol{\delta_\phi},\boldsymbol{\delta_\phi}\rangle}),\\ &\qquad\forall\boldsymbol{\delta_\phi}\in\Phi,\end{aligned}\tag{25}$$

implies

$$y_i\big(\langle\mathbf{w},\phi(\mathbf{x}_i-\boldsymbol{\delta})\rangle+b\big)\geq 1-\xi_i-f_i(\boldsymbol{\delta}),\quad\forall\boldsymbol{\delta}\in\mathbb{R}^n.\tag{26}$$

Notice the condition in Theorem 19 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 20 *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,*

$$\begin{aligned}k(\mathbf{x},\mathbf{x}')&=\hat{k}(d(\mathbf{x},\mathbf{x}')),\quad\forall\mathbf{x},\mathbf{x}'\in\mathbb{R}^n;\\ f_i(\boldsymbol{\delta})&=\hat{f}_i(d(\mathbf{x}_i,\mathbf{x}_i-\boldsymbol{\delta}_i)),\forall\boldsymbol{\delta}\in\mathbb{R}^n.\end{aligned}\tag{27}$$

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

$$h_i(x)=\inf_{y|\exists\mathbf{z}\in\mathbb{R}^n:y=d(\mathbf{x}_i,\mathbf{z}),\,\hat{k}(y)=\hat{k}(0)-x^2/2}\hat{f}_i(y)\tag{28}$$

satisfies Equation (24), and for any $h'(\cdot)$ that satisfies Equation (24), $h'(x)\leq h(x),\forall x\geq 0$ holds. Here, we take $\inf_{y\in\emptyset}\hat{f}_i(y)$ to be $+\infty$.

6 Numerical Simulations

In this section, we report some empirical experiments that were used 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 classifier 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.

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 [30] 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 [31]. 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 captures the setup where noise is skewed toward part of the features, and is more common in practice compared to spherical ones. We repeat 30 such iterations to get the average empirical error of the three different algorithms. Figure 1 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, the comprehensive robust classification has a testing error rate comparable to the 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. The comprehensive robust classification is of interest because it provides a more flexible approach to handle the noise.

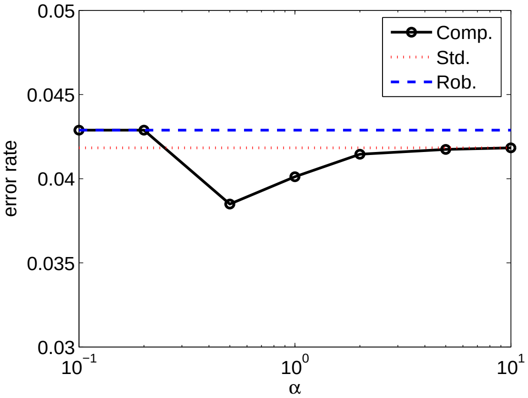

Figure 1: Simulation results for WBC Data.

We run similar simulations on Ionosphere and Sonar data sets from the UCI repository [31]. 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 2 and Figure 3 show the respective simulation results. Similarly to the WBC data set, the comprehensive classification achieves its optimal performance for mid-range α, and is superior to both the standard SVM and the robust SVM.

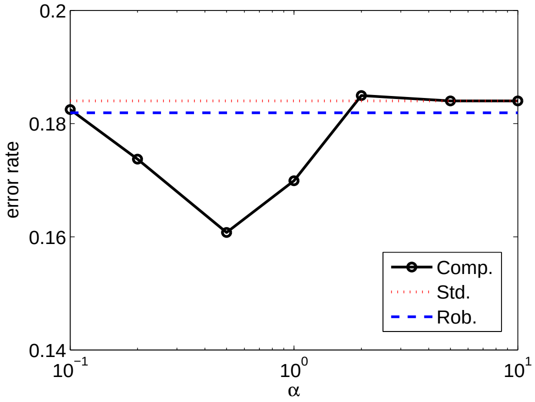

Figure 2: Simulation results for Ionosphere Data.

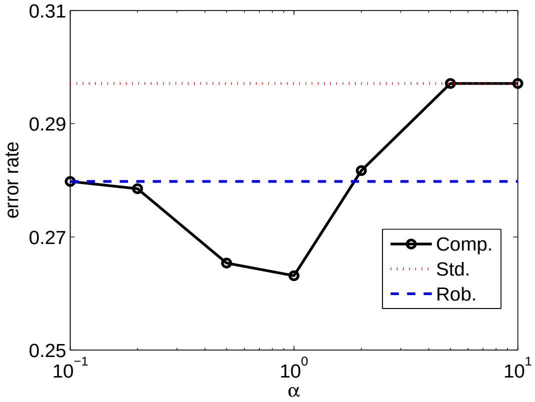

Figure 3: Simulation results for Sonar Data.

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 w.r.t. 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 typically exceedingly 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.

References

[1] 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.

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

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

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

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

[6] 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.

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

[8] 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.

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

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

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

[12] 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, 2004. MIT Press.

[13] 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.

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

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

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

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

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

[19] 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.

[20] 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.

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

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

[23] 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, 2004.

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

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

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

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

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

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

[30] J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999. Special issue on Interior Point Methods (CD supplement with software).

[31] A. Asuncion and D.J. Newman. UCI machine learning repository.