Robustness and Regularization of Support Vector Machines

Huan Xu
ECE, McGill University
Montreal, QC, Canada
xuhuan@cim.mcgill.ca

Constantine Caramanis
ECE, The University of Texas at Austin
Austin, TX, USA
cmcaram@ece.utexas.edu

Shie Mannor
ECE, McGill University
Montreal, QC Canada
shie.mannor@mcgill.ca

Abstract

We consider a robust classification problem and show that standard regularized SVM is a special case of our formulation, providing an explicit link between regularization and robustness. At the same time, the physical connection of noise and robustness suggests the potential for a broad new family of robust classification algorithms. Finally, we show that robustness is a fundamental property of classification algorithms, by re-proving consistency of support vector machines using only robustness arguments (instead of VC dimension or stability).

1 Introduction

Support Vector Machines or SVMs [1, 2] find the hyperplane in the feature space that achieves maximum sample margin in the separable case. When the samples are not separable, a penalty term that approximates the total training-error is considered [3]. It is well known that minimizing the training error itself can lead to poor classification performance for new unlabeled data because of, essentially, overfitting [4]. One of the most popular methods proposed to combat this problem is minimizing a combination of the training-error and a regularization term. 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 [5, 6, 7].

We consider a different setup, assuming that some non-iid (potentially adversarial) disturbance is added to the training samples we observe. We follow a robust optimization approach (e.g., [8, 9]) minimizing the worst possible empirical error under such disturbances. The use of robust optimization in classification is not new (e.g., [10, 11]). Past robust classification models consider only box-type uncertainty sets, which allow the possibility that the data have all been skewed in some non-neutral manner. We develop a new robust classification framework that treats non box-type uncertainty sets, mitigates conservatism, and provides an explicit connection to regularization. Our contributions include:

- We show that the standard regularized SVM classifier is a special case of our robust classification, thus explicitly relating robustness and regularization. This provides an alternative explanation to the success of regularization, and also suggests new physically-motivated ways to construct regularizers.
- Our robust SVM formulation permits finer control of the adversarial disturbance, restricting it to satisfy aggregate constraints across data points, therefore controlling the conservatism.

- We show that the robustness perspective, stemming from a non-iid analysis, can be useful in a standard iid setup, by using it to give a new proof of consistency for standard SVM classification. This result implies that generalization ability is a direct result of robustness to local disturbances, and we can construct learning algorithms that generalize well by robustifying non-consistent algorithms.

We explain here how the explicit equivalence of robustness and regularization we derive differs from previous work, and why it is interesting. While certain equivalence relationships between robustness and regularization have been established for problems outside the machine learning field ([8, 9]), their results do not directly apply to the classification problem. Research on classifier regularization mainly focuses on bounding the complexity of the function class (e.g., [5, 6, 7]). Meanwhile, research on robust classification has not attempted to relate robustness and regularization (e.g., [10, 11, 12]), in part due to the robustness formulations used there.

The connection of robustness and regularization in the SVM context is important for the following reasons. It gives an alternative and potentially powerful explanation of the generalization ability of the regularization term. In the classical machine learning literature, the regularization term bounds the complexity of the class of classifiers. The robust view of regularization regards the testing samples as a perturbed copy of the training samples. We 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 standard 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 Section 3 we use this new view to provide a novel proof of consistency that does not rely on VC-dimension or stability arguments. 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 overly emphasizing the regularizer, and indeed this approach is known to often be too pessimistic ([13]). 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 regularize 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.

Notation: Capital letters and boldface letters are used to denote matrices and column vectors, respectively. For a given norm $\|\cdot\|$, we use $\|\cdot\|^*$ to denote its dual norm. The set of integers from 1 to n is denoted by $[1:n]$.

2 Robust Classification and Regularization

We consider the standard binary classification problem, 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 $h^{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x}\rangle + b)$. For the standard regularized classifier, the parameters $(\mathbf{w}, b)$ are obtained by solving the following convex optimization problem:

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

where $r(\mathbf{w}, b)$ is a regularization term. Previous robust classification work [10, 14] considers the classification problem where the input are subject to (unknown) disturbances $\vec{\boldsymbol{\delta}} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m)$ and essentially solves the following mini-max problem:

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

for a box-type uncertainty set $\mathcal{N}_{\text{box}}$. That is, let $\mathcal{N}_i$ denote the projection of $\mathcal{N}_{\text{box}}$ onto the $\boldsymbol{\delta}_i$ component, then $\mathcal{N}_{\text{box}} = \mathcal{N}_1 \times \cdots \times \mathcal{N}_m$. Effectively, this allows simultaneous worst-case disturbances

across many samples, and leads to overly conservative solutions. The goal of this paper is to obtain a robust formulation where the disturbances $\{\boldsymbol{\delta}_i\}$ may be meaningfully taken to be correlated, i.e., to solve for a non-box-type $\mathcal{N}$:

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

We define explicitly the correlated disturbance (or uncertainty) which we study below.

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 *(CCUS) of $\mathcal{N}_0$, if $\mathcal{N}^- \subseteq \mathcal{N} \subseteq \mathcal{N}^+$. Here*

$$\mathcal{N}^- \triangleq \bigcup_{t=1}^{m} \mathcal{N}_t^-; \quad \mathcal{N}_t^- \triangleq \{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \boldsymbol{\delta}_t \in \mathcal{N}_0;\ \boldsymbol{\delta}_{i \neq t} = \mathbf{0};\}$$

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

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. In particular, $\{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) | \sum_{i=1}^m \|\boldsymbol{\delta}_i\| \leq c\}$ is a CCUS with $\mathcal{N}_0 = \big\{\boldsymbol{\delta} \big| \|\boldsymbol{\delta}\| \leq c\big\}$.

Theorem 1. *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 CCUS with corresponding atomic uncertainty set $\mathcal{N}_0$. Then the following two problems are equivalent:*

$$\min_{\mathbf{w},b} \quad \sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}} \Big\{ r(\mathbf{w},b) + \sum_{i=1}^{m} \max\big[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b),\ 0\big] \Big\}; \tag{3}$$

$$\begin{aligned} \min_{\mathbf{w},b} : \quad & r(\mathbf{w},b) + \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} (\mathbf{w}^\top \boldsymbol{\delta}) + \sum_{i=1}^{m} \xi_i, \\ \text{s.t.} : \quad & \xi_i \geq 1 - [y_i(\langle \mathbf{w},\ \mathbf{x}_i \rangle + b)], \quad i = 1, \ldots, m; \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{4}$$

Proof. We outline the proof. Let $v(\mathbf{w},b) \triangleq \sup_{\boldsymbol{\delta} \in \mathcal{N}_0} (\mathbf{w}^\top \boldsymbol{\delta}) + \sum_{i=1}^m \max\big[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b),\ 0\big]$. It suffices to show that for any $(\hat{\mathbf{w}}, \hat{b}) \in \mathbb{R}^{n+1}$,

$$v(\hat{\mathbf{w}}, \hat{b}) \leq \sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}^-} \sum_{i=1}^{m} \max\big[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\ 0\big]. \tag{5}$$

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

Since the samples $\{\mathbf{x}_i,\ y_i\}_{i=1}^m$ are not separable, there exists $t^* \in [1:m]$ such that $y_{t^*}(\langle \hat{\mathbf{w}}, \mathbf{x}_{t^*} \rangle + \hat{b}) < 0$. With some algebra, we have $\sup_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta_m}) \in \mathcal{N}_{t^*}^-} \sum_{i=1}^m \max\big[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + \hat{b}),\ 0\big] = v(\hat{\mathbf{w}}, \hat{b})$. Since $\mathcal{N}_{t^*}^- \subseteq \mathcal{N}^-$, Inequality (5) follows. Establishing Inequality (6) is standard. ☐

The following corollary thus shows regularized SVM is a special case of robust classification.

Corollary 1. *Let $\mathcal{T} \triangleq \Big\{(\boldsymbol{\delta}_1, \cdots \boldsymbol{\delta}_m) | \sum_{i=1}^m \|\boldsymbol{\delta}_i\| \leq c\Big\}$. If the training samples $\{\mathbf{x}_i, y_i\}_{i=1}^m$ are non-separable, then the following two optimization problems on $(\mathbf{w}, b)$ are equivalent:*

$$\min_{\mathbf{w},b} : \quad \max_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) \in \mathcal{T}_k} \sum_{i=1}^{m} \max\big[1 - y_i\big(\langle \mathbf{w},\ \mathbf{x}_i - \boldsymbol{\delta}_i \rangle + b\big), 0\big], \tag{7}$$

$$\min_{\mathbf{w},b} : \quad c\|\mathbf{w}\|^* + \sum_{i=1}^{m} \max\big[1 - y_i\big(\langle \mathbf{w},\ \mathbf{x}_i \rangle + b\big), 0\big]. \tag{8}$$

This corollary explains the widely known fact that the regularized classifier tends to be robust. It also suggests that the appropriate way to regularize should come from a disturbance-robustness perspective, e.g., by examining the variation of the data and solving the corresponding robust classification problem.

Corollary 1 can be easily generalize to a kernelized version, i.e., a linear classifier in the feature space $\mathcal{H}$ that is defined as a Hilbert space containing the range of some feature mapping $\Phi(\cdot)$.

Corollary 2. *Let* $\mathcal{T}_{\mathcal{H}} \triangleq \Big\{(\boldsymbol{\delta}_1, \cdots \boldsymbol{\delta}_m) | \sum_{i=1}^m \|\boldsymbol{\delta}_i\|_{\mathcal{H}} \leq c\Big\}$. *If* $\{\Phi(\mathbf{x}_i), y_i\}_{i=1}^m$ *are non-separable, then the following two optimization problems on* $(\mathbf{w}, b)$ *are equivalent*

$$\min_{\mathbf{w},b}: \qquad \max_{(\boldsymbol{\delta}_1,\cdots,\boldsymbol{\delta}_m)\in\mathcal{T}_k} \sum_{i=1}^m \max\big[1 - y_i\big(\langle \mathbf{w},\, \Phi(\mathbf{x}_i) - \boldsymbol{\delta}_i\rangle + b\big), 0\big], \tag{9}$$

$$\min_{\mathbf{w},b}: \qquad c\|\mathbf{w}\|_{\mathcal{H}} + \sum_{i=1}^m \max\big[1 - y_i\big(\langle \mathbf{w},\, \Phi(\mathbf{x}_i)\rangle + b\big), 0\big]. \tag{10}$$

Here, $\|\cdot\|_{\mathcal{H}}$ stands for the RKHS norm, which is self-dual. Corollary 2 essentially says that the standard kernelized SVM is implicitly a robust classifier with *disturbance in the feature-space*. Disturbance in the feature-space is less intuitive than disturbance in the sample space. However, the next lemma relates these two different setups: under certain conditions, a classifier that achieves robustness in the feature space (the SVM for example) also achieves robustness in the sample space. The proof is straightforward and omitted.

Lemma 1. *Suppose there exist* $\mathcal{X} \subseteq \mathbb{R}^n$, $\rho > 0$, *and a continuous non-decreasing function* $f : \mathbb{R}^+ \to \mathbb{R}^+$ *satisfying* $f(0) = 0$, *such that*

$$k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2k(\mathbf{x}, \mathbf{x}') \leq f(\|\mathbf{x} - \mathbf{x}'\|_2^2), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \|\mathbf{x} - \mathbf{x}'\|_2 \leq \rho.$$

Then

$$\|\Phi(\hat{\mathbf{x}} + \boldsymbol{\delta}) - \Phi(\hat{\mathbf{x}})\|_{\mathcal{H}} \leq \sqrt{f(\|\boldsymbol{\delta}\|_2^2)}, \quad \forall \|\boldsymbol{\delta}\|_2 \leq \rho,\ \hat{\mathbf{x}}, \hat{\mathbf{x}} + \boldsymbol{\delta} \in \mathcal{X}.$$

3 Consistency of Regularization

In this section we explore a fundamental connection between learning and robustness, by using robustness properties to re-prove the statistical consistency of the linear classifier, and then the kernelized SVM. Indeed, our proof mirrors the consistency proof found in [15], with the key difference that *we replace metric entropy, VC-dimension, and stability used there, with robustness*. In contrast to these standard techniques which often work for a limited range of algorithms, robustness argument works for a much wider range of algorithms and allows a unified approach to show consistency.

Thus far we have considered the setup where the training-samples are corrupted by certain set-inclusive disturbances, and now we turn to the standard statistical learning setup. That is, let $\mathcal{X} \subseteq \mathbb{R}^n$ be bounded, and suppose the training samples $(\mathbf{x}_i, y_i)_{i=1}^{\infty}$ are generated i.i.d. according to an unknown distribution $\mathbb{P}$ supported on $\mathcal{X} \times \{-1, +1\}$. The next theorem shows that our robust classifier setup and equivalently regularized SVM minimizes an asymptotical upper-bound of the expected classification error and hinge loss.

Theorem 2. *Denote* $K \triangleq \max_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x}, \mathbf{x})$. *Suppose there exist* $\rho > 0$ *and a continuous non-decreasing function* $f : \mathbb{R}^+ \to \mathbb{R}^+$ *satisfying* $f(0) = 0$, *such that:*

$$k(\mathbf{x}, \mathbf{x}) + k(\mathbf{x}', \mathbf{x}') - 2k(\mathbf{x}, \mathbf{x}') \leq f(\|\mathbf{x} - \mathbf{x}'\|_2^2), \quad \forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \|\mathbf{x} - \mathbf{x}'\|_2 \leq \rho.$$

Then there exists a random sequence $\{\gamma_{m,c}\}$ *independent of* $\mathbb{P}$ *such that,* $\forall c > 0$, $\lim_{m\to\infty} \gamma_{m,c} = 0$, *almost surely, and* $\forall (\mathbf{w}, b) \in \mathcal{H} \times \mathbb{R}$, *the following bounds on the Bayes loss and the hinge loss hold*

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

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}}\big(\max(1 - y(\langle \mathbf{w},\, \Phi(\mathbf{x})\rangle + b),\, 0)\big) \leq$$

$$\gamma_{m,c}(1 + K\|\mathbf{w}\|_{\mathcal{H}} + |b|) + c\|\mathbf{w}\|_{\mathcal{H}} + \frac{1}{m}\sum_{i=1}^m \max\big[1 - y_i(\langle \mathbf{w},\, \Phi(\mathbf{x}_i)\rangle + b), 0\big].$$

Proof. **Step 1:** We first prove the theorem for a non-kernelized case. We fix a $c > 0$ and drop the subscript c of $\gamma_{m,c}$. 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 m training samples and m testing samples, we use M_m to denote the largest number of pairings.

Lemma 2. *Given $c > 0$, $M_m/m \to 1$ almost surely as $m \to +\infty$, uniformly w.r.t. $\mathbb{P}$.*

Proof. We provide a proof sketch. Partition $\mathcal{X}$ into finite (say T) sets such that if a training sample and a testing sample fall into one set, they form a pairing. This is doable due to finite-dimensionality of the sample space. Let N_i^{tr} and N_i^{te} be the number of training samples and testing samples falling in the i^{th} set, respectively. Thus, $(N_1^{tr}, \cdots, N_T^{tr})$ and $(N_1^{te}, \cdots, N_T^{te})$ are multinomially distributed random vectors following a same distribution. It is straightforward to show that $\sum_{i=1}^{T} \left|N_i^{tr} - N_i^{te}\right|/m \to 0$ with probability one (e.g., Bretagnolle-Huber-Carol inequality), and hence $M_m/m \to 1$ almost surely. Moreover, the convergence rate does not depend on $\mathbb{P}$. □

Now we proceed to prove the theorem. Let $\mathcal{N}_0 = \{\delta \,|\, \|\delta\| \leq c\}$. Given m training samples and m testing samples with M_m sample pairs, for these paired samples, both the total testing error and the total testing hinge-loss are upper bounded by

$$\begin{aligned}
&\max_{(\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m) \in \mathcal{N}_0 \times \cdots \times \mathcal{N}_0} \sum_{i=1}^{m} \max\left[1 - y_i\big(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b\big), 0\right] \\
\leq & cm\|\mathbf{w}\|_2 + \sum_{i=1}^{m} \max\left[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b),\, 0\right].
\end{aligned}$$

Hence the average testing error (including unpaired ones) is upper bounded by

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

Since $\max_{\mathbf{x} \in \mathcal{X}}(1 - y(\langle \mathbf{w}, \mathbf{x}\rangle)) \leq 1 + |b| + K\|\mathbf{w}\|_2$, the average hinge loss is upper bounded by

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

The proof follows since $M_m/m \to 1$ almost surely.

Step 2: Now we generalize the result to a kernelized version. Similarly we lower-bound the number of *sample pairs* in the *feature-space*. The multinomial random variable argument used in the proof of Lemma 2 breaks down, due to possible infinite-dimensionality of the feature space. Nevertheless, we are able to lower bound the number of sample pairs in the feature space by the number of sample pairs in the *sample space*. Define $f^{-1}(\alpha) \triangleq \max\{\beta \geq 0 | f(\beta) \leq \alpha\}$. Since $f(\cdot)$ is continuous, $f^{-1}(\alpha) > 0$ for any $\alpha > 0$. By Lemma 1, if $\mathbf{x}$ and $\mathbf{x}'$ belong to a hyper-cube of length $\min(\rho/\sqrt{n},\, f^{-1}(c^2)/\sqrt{n})$ in the *sample space*, then $\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|_{\mathcal{H}} \leq c$. Hence the number of *sample pairs* in the feature space is lower bounded by the number of pairs of samples that fall in the same hyper-cube in the sample space. We can cover $\mathcal{X}$ with finitely many such hyper-cubes since $f^{-1}(c^2) > 0$. The rest of the proof is identical to Step 1. □

Notice that the condition in Theorem 2 requires that the feature mapping is "smooth" and hence preserves "locality" of the disturbance, i.e., small disturbance in the sample space guarantees the corresponding disturbance in the feature space is also small. This condition is satisfied by most widely used kernels, e.g., homogeneous polynominal kernels, and Gaussian RBF. It is easy to construct non-smooth kernel functions which do not generalize well. For example, consider the following kernel $k(\mathbf{x}, \mathbf{x}') = \mathbf{1}_{(\mathbf{x}=\mathbf{x}')}$, i.e., $k(\mathbf{x}, \mathbf{x}') = 1$ if $\mathbf{x} = \mathbf{x}'$, and zero otherwise. A standard RKHS regularized SVM leads to the a decision function $\operatorname{sign}(\sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b)$, which equals $\operatorname{sign}(b)$ and provides no meaningful prediction if the testing sample $\mathbf{x}$ is not one of the training samples. Hence as m increases, the testing error remains as large as 50% regardless of the tradeoff parameter used in the algorithm, while the training error can be arbitrarily small by fine-tuning the parameter.

Theorem 2 can further lead to the standard consistency notion, i.e., convergence to the Bayes Risk ([15]). The proof in [15] involves a step showing that the expected hinge loss of the minimizer of the regularized *training* hinge loss concentrates around the empirical regularized hinge loss, accomplished using concentration inequalities derived from VC-dimension considerations, and stability considerations. Instead, we can use our robustness-based results of Theorem 2 to replace these approaches. The detailed proof is omitted due to space limits.

4 Concluding Remarks

This work considers the relationship between robustness and regularized SVM classification. In particular, we prove 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. In addition, based on the robustness interpretation of the regularization term, we re-proved the consistency of SVMs without direct appeal to notions of metric entropy, VC-dimension, or stability. Our proof suggests that the ability to handle disturbance is crucial for an algorithm to achieve good generalization ability.

References

[1] P. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin 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] K. Bennett and O. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1(1):23–34, 1992.

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

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

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

[7] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[8] 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.

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

[10] 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.

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

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

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

[14] 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.

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