Journal of Machine Learning Research 10 (2009) 1485-1510 Submitted 3/08; Revised 10/08; Published 7/09

Robustness and Regularization of Support Vector Machines

Huan Xu XUHUAN@CIM.MCGILL.CA
Department of Electrical and Computer Engineering
3480 University Street
McGill University
Montreal, Canada H3A 2A7

Constantine Caramanis CARAMANIS@MAIL.UTEXAS.EDU
Department of Electrical and Computer Engineering
The University of Texas at Austin
1 University Station C0803
Austin, TX, 78712, USA

Shie Mannor* SHIE.MANNOR@MCGILL.CA
Department of Electrical and Computer Engineering
3480 University Street
McGill University
Montreal, Canada H3A 2A7

Editor: Alexander Smola

Abstract

We consider regularized support vector machines (SVMs) and show that they are precisely equivalent to a new robust optimization formulation. We show that this equivalence of robust optimization and regularization has implications for both algorithms, and analysis. In terms of algorithms, the equivalence suggests more general SVM-like algorithms for classification that explicitly build in protection to noise, and at the same time control overfitting. On the analysis front, the equivalence of robustness and regularization provides a robust optimization interpretation for the success of regularized SVMs. We use this new robustness interpretation of SVMs to give a new proof of consistency of (kernelized) SVMs, thus establishing robustness as the *reason* regularized SVMs generalize well.

Keywords: robustness, regularization, generalization, kernel, support vector machine

1. Introduction

Support Vector Machines (SVMs for short) originated in Boser et al. (1992) and can be traced back to as early as Vapnik and Lerner (1963) and Vapnik and Chervonenkis (1974). They continue to be one of the most successful algorithms for classification. SVMs address the classification problem by finding the hyperplane in the feature space that achieves maximum sample margin when the training samples are separable, which leads to minimizing the norm of the classifier. When the samples are not separable, a penalty term that approximates the total training error is considered (Bennett and Mangasarian, 1992; Cortes and Vapnik, 1995). It is well known that minimizing the training error itself can lead to poor classification performance for new unlabeled data; that is, such an approach

∗. Also at the Department of Electrical Engineering, Technion, Israel.

may have poor generalization error because of, essentially, overfitting (Vapnik and Chervonenkis, 1991). A variety of modifications have been proposed to handle this, one of the most popular methods being that of minimizing a combination of the training-error and a regularization term. The latter is typically chosen as a norm of the classifier. The resulting regularized classifier performs better on new data. This phenomenon is often interpreted from a statistical learning theory view: the regularization term restricts the complexity of the classifier, hence the deviation of the testing error and the training error is controlled (see Smola et al., 1998; Evgeniou et al., 2000; Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002; Bartlett et al., 2005, and references therein).

In this paper we consider a different setup, assuming that the training data are generated by the true underlying distribution, but some non-i.i.d. (potentially adversarial) disturbance is then added to the samples we observe. We follow a robust optimization (see El Ghaoui and Lebret, 1997; Ben-Tal and Nemirovski, 1999; Bertsimas and Sim, 2004, and references therein) approach, that is, minimizing the worst possible empirical error under such disturbances. The use of robust optimization in classification is not new (e.g., Shivaswamy et al., 2006; Bhattacharyya et al., 2004b; Lanckriet et al., 2003), in which *box-type* uncertainty sets were considered. Moreover, there has not been an explicit connection to the regularized classifier, although at a high-level it is known that regularization and robust optimization are related (e.g., El Ghaoui and Lebret, 1997; Anthony and Bartlett, 1999). The main contribution in this paper is solving the robust classification problem for a class of non-box-typed uncertainty sets, and providing a linkage between robust classification and the standard regularization scheme of SVMs. In particular, our contributions include the following:

- We solve the robust SVM formulation for a class of non-box-type uncertainty sets. This permits finer control of the adversarial disturbance, restricting it to satisfy aggregate constraints across data points, therefore reducing the possibility of highly correlated disturbance.

- 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 regularization terms.

- We relate our robust formulation to several probabilistic formulations. We consider a chance-constrained classifier (that is, a classifier with probabilistic constraints on misclassification) and show that our robust formulation can approximate it far less conservatively than previous robust formulations could possibly do. We also consider a Bayesian setup, and show that this can be used to provide a principled means of selecting the regularization coefficient without cross-validation.

- We show that the robustness perspective, stemming from a non-i.i.d. analysis, can be useful in the standard learning (i.i.d.) setup, by using it to prove consistency for standard SVM classification, *without using VC-dimension or stability arguments*. This result implies that generalization ability is a direct result of robustness to local disturbances; it therefore suggests a new justification for good performance, and consequently allows us to construct learning algorithms that generalize well by robustifying non-consistent algorithms.

1.1 Robustness and Regularization

We comment here on the explicit equivalence of robustness and regularization. We briefly explain how this observation is different from previous work and why it is interesting. Previous works on robust classification (e.g., Lanckriet et al., 2003; Bhattacharyya et al., 2004a,b; Shivaswamy et al., 2006; Trafalis and Gilbert, 2007) consider robustifying *regularized* classifications.[1] That is, they propose modifications to standard regularized classifications so that the new formulations are robust to input uncertainty. Furthermore, box-type uncertainty—a setup where the joint uncertainty is the Cartesian product of uncertainty in each input (see Section 2 for detailed formulation)—is considered, which leads to penalty terms on each *constraint* of the resulting formulation. The objective of these works was not to relate robustness and the standard regularization term that appears in the *objective function*. Indeed, research on classifier regularization mainly considers its effect on bounding the complexity of the function class (e.g., Smola et al., 1998; Evgeniou et al., 2000; Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002; Bartlett et al., 2005). Thus, although certain equivalence relationships between robustness and regularization have been established for problems other than classification (El Ghaoui and Lebret, 1997; Ben-Tal and Nemirovski, 1999; Bishop, 1995; Xu et al., 2009), the explicit equivalence between robustness and regularization in the SVM setup is novel.

The connection of robustness and regularization in the SVM context is important for the following reasons. First, it gives an alternative and potentially powerful explanation of the generalization ability of the regularization term. In the standard machine learning view, 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. Therefore, 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.

Second, this connection suggests novel approaches to designing good classification algorithms, in particular, designing the regularization term. In the PAC structural-risk minimization approach, regularization is chosen to minimize a bound on the generalization error based on the training error and a complexity term. This approach is known to often be too pessimistic (Kearns et al., 1997), especially for problems with more structure. 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 known that normalizing the samples so that the variance among all features is roughly the same (a process commonly used to eliminate the scaling freedom of individual features) often leads to good generalization performance. From the robustness perspective, this has the interpretation that the noise is anisotropic (ellipsoidal) rather than spherical, and hence an appropriate robustification must be designed to fit this anisotropy.

We also show that using the robust optimization viewpoint, we obtain some probabilistic results that go beyond the PAC setup. In Section 3 we bound the probability that a noisy training sample is correctly labeled. Such a bound considers the behavior of *corrupted* samples and is hence different from the known PAC bounds. This is helpful when the training samples and the testing samples are

1. Lanckriet et al. (2003) is perhaps the only exception, where a regularization term is added to the covariance estimation rather than to the objective function.

drawn from different distributions, or some adversary manipulates the samples to prevent them from being correctly labeled (e.g., spam senders change their patterns from time to time to avoid being labeled and filtered). Finally, this connection of robustification and regularization also provides us with new proof techniques as well (see Section 5).

We need to point out that there are several different definitions of robustness in the literature. In this paper, as well as the aforementioned robust classification papers, robustness is mainly understood from a Robust Optimization (RO) perspective, where a min-max optimization is performed over all possible disturbances. An alternative interpretation of robustness stems from the rich literature on robust statistics (e.g., Huber, 1981; Hampel et al., 1986; Rousseeuw and Leroy, 1987; Maronna et al., 2006), which studies how an estimator or algorithm behaves under a small perturbation of the statistics model. For example, the *influence function* approach, proposed in Hampel (1974) and Hampel et al. (1986), measures the impact of an infinitesimal amount of contamination of the original distribution on the quantity of interest. Based on this notion of robustness, Christmann and Steinwart (2004) showed that many kernel classification algorithms, including SVM, are robust in the sense of having a finite Influence Function. A similar result for regression algorithms is shown in Christmann and Steinwart (2007) for smooth loss functions, and in Christmann and Van Messem (2008) for non-smooth loss functions where a relaxed version of the Influence Function is applied. In the machine learning literature, another widely used notion closely related to robustness is the *stability*, where an algorithm is required to be robust (in the sense that the output function does not change significantly) under a specific perturbation: deleting one sample from the training set. It is now well known that a stable algorithm such as SVM has desirable generalization properties, and is statistically consistent under mild technical conditions; see for example Bousquet and Elisseeff (2002), Kutin and Niyogi (2002), Poggio et al. (2004) and Mukherjee et al. (2006) for details. One main difference between RO and other robustness notions is that the former is constructive rather than analytical. That is, in contrast to robust statistics or the stability approach that measures the robustness of a *given* algorithm, RO can *robustify* an algorithm: it converts a given algorithm to a robust one. For example, as we show in this paper, the RO version of a naive empirical-error minimization is the well known SVM. As a constructive process, the RO approach also leads to additional flexibility in algorithm design, especially when the nature of the perturbation is known or can be well estimated.

1.2 Structure of the Paper

This paper is organized as follows. In Section 2 we investigate the correlated disturbance case, and show the equivalence between the robust classification and the regularization process. We develop the connections to probabilistic formulations in Section 3. The kernelized version is investigated in Section 4. Finally, in Section 5, we consider the standard statistical learning setup where all samples are i.i.d. draws and provide a novel proof of consistency of SVM based on robustness analysis. The analysis shows that duplicate copies of iid draws tend to be "similar" to each other in the sense that with high probability the total difference is small, and hence robustification that aims to control performance loss for small perturbations can help mitigate overfitting even though no explicit perturbation exists.

1.3 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, that is, $\|\mathbf{z}\|^* \triangleq \sup\{\mathbf{z}^\top \mathbf{x} | \|\mathbf{x}\| \leq 1\}$. 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 δ_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

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

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

where $r(\mathbf{w},b)$ is a regularization term. This is equivalent to

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

Previous robust classification work (Shivaswamy et al., 2006; Bhattacharyya et al., 2004a,b; Bhattacharyya, 2004; Trafalis and Gilbert, 2007) considers the classification problem where the input are subject to (unknown) disturbances $\vec{\delta} = (\delta_1, \ldots, \delta_m)$ and essentially solves the following min-max problem:

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

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 δ_i component, then $\mathcal{N}_{\text{box}} = \mathcal{N}_1 \times \cdots \times \mathcal{N}_m$ (note that $\mathcal{N}_i$ need not be a "box"). 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 $\{\delta_i\}$ may be meaningfully taken to be correlated, that is, to solve for a non-box-type $\mathcal{N}$:

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

We briefly explain here the four reasons that motivate this "robust to perturbation" setup and in particular the min-max form of (1) and (2). First, it can explicitly incorporate prior problem knowledge of local invariance (e.g., Teo et al., 2008). For example, in vision tasks, a desirable classifier should provide a consistent answer if an input image slightly changes. Second, there are situations where

some adversarial opponents (e.g., spam senders) will manipulate the testing samples to avoid being correctly classified, and the robustness toward such manipulation should be taken into consideration in the training process (e.g., Globerson and Roweis, 2006). Or alternatively, the training samples and the testing samples can be obtained from different processes and hence the standard i.i.d. assumption is violated (e.g., Bi and Zhang, 2004). For example in real-time applications, the newly generated samples are often less accurate due to time constraints. Finally, formulations based on chance-constraints (e.g., Bhattacharyya et al., 2004b; Shivaswamy et al., 2006) are mathematically equivalent to such a min-max formulation.

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

Definition 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) *For any* $\mathbf{w}_0 \in \mathbb{R}^n$: $\sup_{\delta \in \mathcal{N}_0} [\mathbf{w}_0^\top \delta] = \sup_{\delta' \in \mathcal{N}_0} [-\mathbf{w}_0^\top \delta'] < +\infty$.

We use "sup" here because the maximal value is not necessary attained since $\mathcal{N}_0$ may not be a closed set. The second condition of Atomic Uncertainty set basically says that the uncertainty set is bounded and symmetric. In particular, all norm balls and ellipsoids centered at the origin are atomic uncertainty sets, while an arbitrary polytope might not be an atomic uncertainty set.

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

$$\mathcal{N}^- \subseteq \mathcal{N} \subseteq \mathcal{N}^+,$$

$$\textit{where:} \quad \mathcal{N}^- \triangleq \bigcup_{t=1}^{m} \mathcal{N}_t^-; \qquad \mathcal{N}_t^- \triangleq \{(\delta_1, \cdots, \delta_m) | \delta_t \in \mathcal{N}_0;\ \delta_{i\neq t} = \mathbf{0}\}.$$

$$\mathcal{N}^+ \triangleq \{(\alpha_1\delta_1, \cdots, \alpha_m\delta_m) | \sum_{i=1}^{m} \alpha_i = 1;\ \alpha_i \geq 0,\ \delta_i \in \mathcal{N}_0, i = 1, \cdots, m\}.$$

The Sublinear Aggregated 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

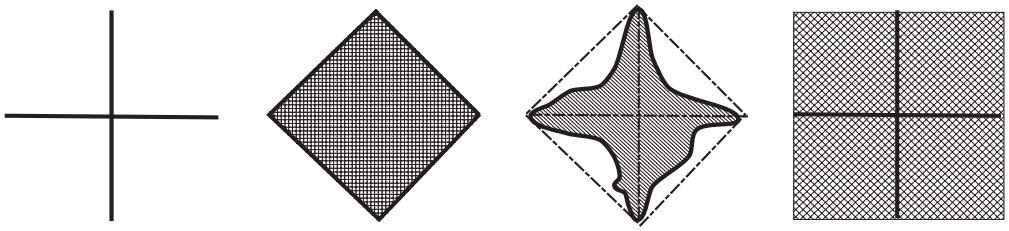
(1) $\{(\delta_1, \cdots, \delta_m) | \sum_{i=1}^{m} \|\delta_i\| \leq c\}$;

(2) $\{(\delta_1, \cdots, \delta_m) | \exists t \in [1:m];\ \|\delta_t\| \leq c;\ \delta_i = \mathbf{0}, \forall i \neq t\}$;

(3) $\{(\delta_1, \cdots, \delta_m) | \sum_{i=1}^{m} \sqrt{c\|\delta_i\|} \leq c\}$.

All these examples have the same atomic uncertainty set $\mathcal{N}_0 = \left\{\delta \middle| \|\delta\| \leq c\right\}$. Figure 1 provides an illustration of a sublinear aggregated uncertainty set for $n = 1$ and $m = 2$, that is, the training set consists of two univariate samples.

The following theorem is the main result of this section, which reveals that standard norm regularized SVM is the solution of a (non-regularized) robust optimization. It is a special case of Proposition 4 by taking $\mathcal{N}_0$ as the dual-norm ball $\{\delta | \|\delta\|^* \leq c\}$ for an arbitrary norm $\|\cdot\|$ and $r(\mathbf{w}, b) \equiv 0$.



a. $\mathcal{N}^-$ b. $\mathcal{N}^+$ c. $\mathcal{N}$ d. Box uncertainty

Figure 1: Illustration of a Sublinear Aggregated Uncertainty Set $\mathcal{N}$.

Theorem 3 *Let* $\mathcal{T} \triangleq \left\{(\delta_1, \cdots \delta_m) | \sum_{i=1}^m \|\delta_i\|^* \leq c\right\}$. *Suppose that the training sample* $\{\mathbf{x}_i, y_i\}_{i=1}^m$ *are non-separable. Then the following two optimization problems on* $(\mathbf{w}, b)$ *are equivalent*[2]

$$\begin{array}{ll} \min: & \max\limits_{(\delta_1,\cdots,\delta_m)\in\mathcal{T}} \sum\limits_{i=1}^m \max\left[1 - y_i\big(\langle \mathbf{w}, \mathbf{x}_i - \delta_i\rangle + b\big), 0\right], \\ \min: & c\|\mathbf{w}\| + \sum\limits_{i=1}^m \max\left[1 - y_i\big(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\big), 0\right]. \end{array} \tag{3}$$

Proposition 4 *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 Sublinear Aggregated Uncertainty set with corresponding atomic uncertainty set* $\mathcal{N}_0$. *Then the following min-max problem*

$$\min_{\mathbf{w},b} \sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}} \left\{ r(\mathbf{w}, b) + \sum_{i=1}^m \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i - \delta_i\rangle + b), 0\right]\right\} \tag{4}$$

is equivalent to the following optimization problem on $\mathbf{w}, b, \xi$*:*

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

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

Proof Define:

$$v(\mathbf{w}, b) \quad \triangleq \quad \sup_{\delta\in\mathcal{N}_0} (\mathbf{w}^\top \delta) + \sum_{i=1}^m \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b), 0\right].$$

2. The optimization equivalence for the linear case was observed independently by Bertsimas and Fertis (2008).

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

$$\begin{aligned}
&\sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}^-} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \delta_i\rangle + \hat{b}), 0\right] \\
\leq &\sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \delta_i\rangle + \hat{b}), 0\right] \\
\leq &\sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}^+} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \delta_i\rangle + \hat{b}), 0\right].
\end{aligned}$$

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

Step 1: We prove that

$$v(\hat{\mathbf{w}}, \hat{b}) \leq \sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}^-} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \delta_i\rangle + \hat{b}), 0\right]. \tag{6}$$

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

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

Hence,

$$\begin{aligned}
&\sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}_t^-} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \delta_i\rangle + \hat{b}), 0\right] \\
&= \sum_{i\neq t} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle + \hat{b}), 0\right] + \sup_{\delta_t\in\mathcal{N}_0} \max\left[1 - y_t(\langle \hat{\mathbf{w}}, \mathbf{x}_t - \delta_t\rangle + \hat{b}), 0\right] \\
&= \sum_{i\neq t} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle + \hat{b}), 0\right] + \max\left[1 - y_t(\langle \hat{\mathbf{w}}, \mathbf{x}_t\rangle + \hat{b}) + \sup_{\delta_t\in\mathcal{N}_0} (y_t \hat{\mathbf{w}}^\top \delta_t), 0\right] \\
&= \sum_{i\neq t} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle + \hat{b}), 0\right] + \max\left[1 - y_t(\langle \hat{\mathbf{w}}, \mathbf{x}_t\rangle + \hat{b}), 0\right] + \sup_{\delta_t\in\mathcal{N}_0} (y_t \hat{\mathbf{w}}^\top \delta_t) \\
&= \sup_{\delta\in\mathcal{N}_0} (\hat{\mathbf{w}}^\top \delta) + \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i\rangle + \hat{b}), 0\right] = v(\hat{\mathbf{w}}, \hat{b}).
\end{aligned}$$

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

Step 2: Next we prove that

$$\sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}^+} \sum_{i=1}^{m} \max\left[1 - y_i(\langle \hat{\mathbf{w}}, \mathbf{x}_i - \delta_i\rangle + \hat{b}), 0\right] \leq v(\hat{\mathbf{w}}, \hat{b}). \tag{8}$$

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

$$\begin{aligned}
&\sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}^+}\sum_{i=1}^m \max\left[1-y_i(\langle\hat{\mathbf{w}},\mathbf{x}_i-\delta_i\rangle+\hat{b}),0\right]\\
=&\sup_{\sum_{i=1}^m\alpha_i=1;\alpha_i\geq 0;\hat{\delta}_i\in\mathcal{N}_0}\sum_{i=1}^m \max\left[1-y_i(\langle\hat{\mathbf{w}},\mathbf{x}_i-\alpha_i\hat{\delta}_i\rangle+\hat{b}),0\right] \qquad (9)\\
=&\sup_{\sum_{i=1}^m\alpha_i=1;\alpha_i\geq 0;}\sum_{i=1}^m \max\Big[\sup_{\hat{\delta}_i\in\mathcal{N}_0}\left(1-y_i(\langle\hat{\mathbf{w}},\mathbf{x}_i-\alpha_i\hat{\delta}_i\rangle+\hat{b})\right),0\Big].
\end{aligned}$$

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

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

Therefore, Equation (9) is upper bounded by

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

hence Inequality (8) holds.

Step 3: Combining the two steps and adding $r(\mathbf{w},b)$ on both sides leads to: $\forall(\mathbf{w},b)\in\mathbb{R}^{n+1}$,

$$\sup_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}}\sum_{i=1}^m\max\left[1-y_i(\langle\mathbf{w},\mathbf{x}_i-\delta_i\rangle+b),0\right]+r(\mathbf{w},b)=v(\mathbf{w},b)+r(\mathbf{w},\mathbf{b}).$$

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

Before concluding this section we briefly comment on the meaning of Theorem 3 and Proposition 4. On one hand, they explain the widely known fact that the regularized classifier tends to be more robust (see for example, Christmann and Steinwart, 2004, 2007; Christmann and Van Messem, 2008; Trafalis and Gilbert, 2007). 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, particularly if some a-priori knowledge of the data-variation is known. For example, consider an image

classification problem. Suppose it is known that these pictures are taken under significantly varying background light. Therefore, for a given sample (picture), the perturbation on each feature (pixel) is large. However, the perturbations across different features are almost identical since they are under the same background light. This can be represented by the following Atomic uncertainty set

$$\mathcal{N}_0 = \{\delta | \|\delta\|_2 \le c_1,\ \|\delta - (\frac{1}{n}\sum_{t=1}^{n}\delta_t)\mathbf{1}\|_2 \le c_2\},$$

where $c_2 \ll c_1$. By Proposition 4, this leads to the following regularization term

$$\begin{aligned} f(\mathbf{w}) = \ & \max : \mathbf{w}^\top \delta \\ & \text{s.t.} \ \|\delta\|_2 \le c_1 \\ & \quad \|(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\delta\|_2 \le c_2. \end{aligned}$$

Notice this is a second order cone programming which has a dual form

$$\begin{aligned} & \min : \ c_1 v_1 + c_2 v_2 \\ & \text{s.t.} \ \mathbf{u}_1 + (I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\mathbf{u}_2 = \mathbf{w} \\ & \quad \|\mathbf{u}_i\|_2 \le v_i,\ i = 1, 2. \end{aligned}$$

Substituting it to (5), the resulting classification problem is a second order cone program, which can be efficiently solved (Boyd and Vandenberghe, 2004).

3. Probabilistic Interpretations

Although Problem (4) is formulated without any probabilistic assumptions, in this section, we briefly explain two approaches to construct the uncertainty set and equivalently tune the regularization parameter c based on probabilistic information.

The first approach is to use Problem (4) to approximate an upper bound for a chance-constrained classifier. Suppose the disturbance $(\delta_1^r, \cdots \delta_m^r)$ follows a joint probability measure μ. Then the chance-constrained classifier is given by the following minimization problem given a confidence level $\eta \in [0, 1]$,

$$\begin{aligned} \min_{\mathbf{w},b,l} : \quad & l \\ \text{s.t.} : \quad & \mu\Big\{\sum_{i=1}^{m} \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i - \delta_i^r\rangle + b), 0\right] \le l\Big\} \ge 1 - \eta. \end{aligned} \tag{10}$$

The formulations in Shivaswamy et al. (2006), Lanckriet et al. (2003) and Bhattacharyya et al. (2004a) assume uncorrelated noise and require all constraints to be satisfied with high probability *simultaneously*. They find a vector $[\xi_1, \cdots, \xi_m]^\top$ where each ξ_i is the η-quantile of the hinge-loss for sample $\mathbf{x}_i^r$. In contrast, our formulation above minimizes the η-quantile of the average (or equivalently the sum of) empirical error. When controlling this average quantity is of more interest, the box-type noise formulation will be overly conservative.

Problem (10) is generally intractable. However, we can approximate it as follows. Let

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

Notice that c^* is easily simulated given μ. Then for any $(\mathbf{w}, b)$, with probability no less than $1-\eta$, the following holds,

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

Thus (10) is upper bounded by (3) 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, that is, the box uncertainty set, would lead to considerably more pessimistic approximations of the chance constraint.

The second approach considers a Bayesian setup. Suppose the total disturbance $c^r \triangleq \sum_{i=1}^m \|\delta_i^r\|^*$ follows a prior distribution $\rho(\cdot)$. This can model for example the case that the training sample set is a mixture of several data sets where the disturbance magnitude of each set is known. Such a setup leads to the following classifier which minimizes the Bayesian (robust) error:

$$\min_{\mathbf{w},b}: \quad \int \left\{ \max_{\sum \|\delta_i\|^* \leq c} \sum_{i=1}^{m} \max\left[1 - y_i\big(\langle \mathbf{w}, \mathbf{x}_i - \delta_i \rangle + b\big), 0\right] \right\} d\rho(c). \tag{11}$$

By Theorem 3, the Bayes classifier (11) is equivalent to

$$\min_{\mathbf{w},b}: \quad \int \left\{ c\|\mathbf{w}\| + \sum_{i=1}^{m} \max\left[1 - y_i\big(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\big), 0\right] \right\} d\rho(c),$$

which can be further simplified as

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

where $\bar{c} \triangleq \int c\, d\rho(c)$. This thus provides us a justifiable parameter tuning method different from cross validation: simply using the expected value of c^r. We note that it is the equivalence of Theorem 3 that makes this possible, since it is difficult to imagine a setting where one would have a prior on regularization coefficients.

4. Kernelization

The previous results can be easily generalized to the kernelized setting, which we discuss in detail in this section. In particular, similar to the linear classification case, we give a new interpretation of the standard kernelized SVM as the min-max empirical hinge-loss solution, where the disturbance is assumed to lie in the feature space. We then relate this to the (more intuitively appealing) setup where the disturbance lies in the sample space. We use this relationship in Section 5 to prove a consistency result for kernelized SVMs.

The kernelized SVM formulation considers a linear classifier in the feature space $\mathcal{H}$, a Hilbert space containing the range of some feature mapping $\Phi(\cdot)$. The standard formulation is as follows,

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

It has been proved in Schölkopf and Smola (2002) that if we take $f(\langle \mathbf{w}, \mathbf{w} \rangle)$ for some increasing function $f(\cdot)$ as the regularization term $r(\mathbf{w}, b)$, then the optimal solution has a representation $\mathbf{w}^* = \sum_{i=1}^m \alpha_i \Phi(\mathbf{x}_i)$, which can further be solved without knowing explicitly the feature mapping, but by evaluating a kernel function $k(\mathbf{x}, \mathbf{x}') \triangleq \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$ only. This is the well-known "kernel trick".

The definitions of Atomic Uncertainty Set and Sublinear Aggregated Uncertainty Set in the feature space are identical to Definition 1 and 2, with $\mathbb{R}^n$ replaced by $\mathcal{H}$. The following theorem is a feature-space counterpart of Proposition 4. The proof follows from a similar argument to Proposition 4, that is, for any fixed $(\mathbf{w}, b)$ the worst-case empirical error equals the empirical error plus a penalty term $\sup_{\delta \in \mathcal{N}_0} \big(\langle \mathbf{w}, \delta \rangle \big)$, and hence the details are omitted.

Theorem 5 *Assume $\{\Phi(\mathbf{x}_i), y_i\}_{i=1}^m$ are not linearly separable, $r(\cdot) : \mathcal{H} \times \mathbb{R} \to \mathbb{R}$ is an arbitrary function, $\mathcal{N} \subseteq \mathcal{H}^m$ is a Sublinear Aggregated Uncertainty set with corresponding atomic uncertainty set $\mathcal{N}_0 \subseteq \mathcal{H}$. Then the following min-max problem*

$$\min_{\mathbf{w}, b} \sup_{(\delta_1, \cdots, \delta_m) \in \mathcal{N}} \left\{ r(\mathbf{w}, b) + \sum_{i=1}^m \max \left[1 - y_i (\langle \mathbf{w}, \Phi(\mathbf{x}_i) - \delta_i \rangle + b), 0 \right] \right\}$$

is equivalent to

$$\begin{array}{rll} \min: & r(\mathbf{w}, b) + \sup\limits_{\delta \in \mathcal{N}_0} \left(\langle \mathbf{w}, \delta \rangle \right) + \sum\limits_{i=1}^m \xi_i, & \\ \text{s.t.}: & \xi_i \geq 1 - y_i \big(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b \big), & i = 1, \cdots, m; \\ & \xi_i \geq 0, & i = 1, \cdots, m. \end{array} \tag{12}$$

Furthermore, the minimization of Problem (12) is attainable when $r(\cdot, \cdot)$ is lower semi-continuous.

For some widely used feature mappings (e.g., RKHS of a Gaussian kernel), $\{\Phi(\mathbf{x}_i), y_i\}_{i=1}^m$ are always separable. In this case, the worst-case empirical error may not be equal to the empirical error plus a penalty term $\sup_{\delta \in \mathcal{N}_0} \big(\langle \mathbf{w}, \delta \rangle \big)$. However, it is easy to show that for any $(\mathbf{w}, b)$, the latter is an upper bound of the former.

The next corollary is the feature-space counterpart of Theorem 3, where $\| \cdot \|_{\mathcal{H}}$ stands for the RKHS norm, that is, for $\mathbf{z} \in \mathcal{H}$, $\|\mathbf{z}\|_{\mathcal{H}} = \sqrt{\langle \mathbf{z}, \mathbf{z} \rangle}$. Noticing that the RKHS norm is self dual, we find that the proof is identical to that of Theorem 3, and hence omit it.

Corollary 6 *Let $\mathcal{T}_{\mathcal{H}} \triangleq \left\{ (\delta_1, \cdots \delta_m) \middle| \sum_{i=1}^m \|\delta_i\|_{\mathcal{H}} \leq c \right\}$. 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*

$$\begin{array}{ll} \min: & \max\limits_{(\delta_1, \cdots, \delta_m) \in \mathcal{T}_{\mathcal{H}}} \sum\limits_{i=1}^m \max \left[1 - y_i \big(\langle \mathbf{w}, \Phi(\mathbf{x}_i) - \delta_i \rangle + b \big), 0 \right], \\ \min: & c \|\mathbf{w}\|_{\mathcal{H}} + \sum\limits_{i=1}^m \max \left[1 - y_i \big(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b \big), 0 \right]. \end{array} \tag{13}$$

Equation (13) is a variant form of the standard SVM that has a squared RKHS norm regularization term, and it can be shown that the two formulations are equivalent up to changing of tradeoff parameter c, since both the empirical hinge-loss and the RKHS norm are convex. Therefore, Corollary 6

essentially means that the standard kernelized SVM is implicitly a robust classifier (without regularization) with disturbance in the feature-space, and the sum of the magnitude of the disturbance is bounded.

Disturbance in the feature-space is less intuitive than disturbance in the sample space, and the next lemma relates these two different notions.

Lemma 7 *Suppose there exists $\mathcal{X} \subseteq \mathbb{R}^n$, $\rho > 0$, 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

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

In the appendix, we prove a result that provides a tighter relationship between disturbance in the feature space and disturbance in the sample space, for RBF kernels.

Proof Expanding the RKHS norm yields

$$\begin{aligned}
&\|\Phi(\hat{\mathbf{x}}+\delta) - \Phi(\hat{\mathbf{x}})\|_{\mathcal{H}} \\
=&\sqrt{\langle \Phi(\hat{\mathbf{x}}+\delta) - \Phi(\hat{\mathbf{x}}), \Phi(\hat{\mathbf{x}}+\delta) - \Phi(\hat{\mathbf{x}})\rangle} \\
=&\sqrt{\langle \Phi(\hat{\mathbf{x}}+\delta), \Phi(\hat{\mathbf{x}}+\delta)\rangle + \langle \Phi(\hat{\mathbf{x}}), \Phi(\hat{\mathbf{x}})\rangle - 2\langle \Phi(\hat{\mathbf{x}}+\delta), \Phi(\hat{\mathbf{x}})\rangle} \\
=&\sqrt{k\big(\hat{\mathbf{x}}+\delta, \hat{\mathbf{x}}+\delta\big) + k\big(\hat{\mathbf{x}}, \hat{\mathbf{x}}\big) - 2k\big(\hat{\mathbf{x}}+\delta,\ \hat{\mathbf{x}}\big)} \\
\leq&\sqrt{f(\|\hat{\mathbf{x}}+\delta-\hat{\mathbf{x}}\|_2^2)} = \sqrt{f(\|\delta\|_2^2)},
\end{aligned}$$

where the inequality follows from the assumption. ■

Lemma 7 essentially says that under certain conditions, robustness in the feature space is a stronger requirement that robustness in the sample space. Therefore, a classifier that achieves robustness in the feature space (the SVM for example) also achieves robustness in the sample space. Notice that the condition of Lemma 7 is rather weak. In particular, it holds for any continuous $k(\cdot,\cdot)$ and bounded $\mathcal{X}$.

In the next section we consider a more foundational property of robustness in the sample space: we show that a classifier that is robust in the sample space is asymptotically consistent. As a consequence of this result for linear classifiers, the above results imply the consistency for a broad class of kernelized SVMs.

5. 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 Steinwart (2005), with the key difference that *we replace metric entropy, VC-dimension, and stability conditions used there, with a robustness condition.*

Thus far we have considered the setup where the training-samples are corrupted by certain set-inclusive disturbances. We now turn to the standard statistical learning setup, by assuming that all

training samples and testing samples are generated i.i.d. according to a (unknown) probability $\mathbb{P}$, that is, there does not exist explicit disturbance.

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 by $\mathcal{X} \times \{-1, +1\}$. The next theorem shows that our robust classifier setup and equivalently regularized SVM asymptotically minimizes an upper-bound of the expected classification error and hinge loss.

Theorem 8 *Denote* $K \triangleq \max_{x \in \mathcal{X}} \|x\|_2$. *Then there exists a random sequence* $\{\gamma_{m,c}\}$ *such that:*

1. $\forall c > 0$, $\lim_{m\to\infty} \gamma_{m,c} = 0$ *almost surely, and the convergence is uniform in* $\mathbb{P}$*;*

2. *the following bounds on the Bayes loss and the hinge loss hold uniformly for all* $(\mathbf{w}, b)$*:*

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

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

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

Proof We briefly explain the basic idea of the proof before going to the technical details. We consider the testing sample set as a perturbed copy of the training sample set, and measure the magnitude of the perturbation. For testing samples that have "small" perturbations, $c\|\mathbf{w}\|_2 + \frac{1}{m}\sum_{i=1}^{m} \max\left[1 - y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b), 0\right]$ upper-bounds their total loss by Theorem 3. Therefore, we only need to show that the ratio of testing samples having "large" perturbations diminishes to prove the theorem.

Now we present the detailed proof. Given a $c > 0$, we call a testing sample $(\mathbf{x}', y')$ and a training sample $(\mathbf{x}, y)$ 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 $M_{m,c}$ to denote the largest number of pairings. To prove this theorem, we need to establish the following lemma.

Lemma 9 *Given a* $c > 0$, $M_{m,c}/m \to 1$ *almost surely as* $m \to +\infty$, *uniformly w.r.t.* $\mathbb{P}$.

Proof We make a partition of $\mathcal{X} \times \{-1, +1\} = \bigcup_{t=1}^{T_c} \mathcal{X}_t$ such that $\mathcal{X}_t$ 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 $[\alpha_1, \alpha_1 + c/\sqrt{n}) \times [\alpha_2, \alpha_2 + c/\sqrt{n}) \cdots \times [\alpha_n, \alpha_n + c/\sqrt{n}) \times \{-1\}$ (recall n is the dimension of $\mathcal{X}$). 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 $\mathcal{X}_t$, they can form a pairing.

Let N_t^{tr} and N_t^{te} be the number of training samples and testing samples falling in the t^{th} set, respectively. Thus, $(N_1^{tr}, \cdots, N_{T_c}^{tr})$ and $(N_1^{te}, \cdots, N_{T_c}^{te})$ are multinomially distributed random vectors following a same distribution. Notice that for a multinomially distributed random vector $(N_1, \cdots, N_k)$ with parameter m and $(p_1, \cdots, p_k)$, the following holds (Bretegnolle-Huber-Carol inequality, see for example Proposition A6.6 of van der Vaart and Wellner, 2000). For any $\lambda > 0$,

$$\mathbb{P}\Big(\sum_{i=1}^{k} \big|N_i - mp_i\big|) \geq 2\sqrt{m}\lambda\Big) \leq 2^k \exp(-2\lambda^2).$$

Hence we have

$$\begin{aligned} & \mathbb{P}\Big(\sum_{t=1}^{T_c}\big|N_t^{tr}-N_t^{te}\big|\geq 4\sqrt{m}\lambda\Big)\leq 2^{T_c+1}\exp(-2\lambda^2), \\ \Longrightarrow \quad & \mathbb{P}\Big(\frac{1}{m}\sum_{t=1}^{T_c}\big|N_t^{tr}-N_t^{te}\big|\geq \lambda\Big)\leq 2^{T_c+1}\exp(\frac{-m\lambda^2}{8}), \\ \Longrightarrow \quad & \mathbb{P}\Big(M_{m,c}/m\leq 1-\lambda\Big)\leq 2^{T_c+1}\exp(\frac{-m\lambda^2}{8}). \end{aligned} \tag{14}$$

Observe that $\sum_{m=1}^{\infty}2^{T_c+1}\exp(\frac{-m\lambda^2}{8})<+\infty$, hence by the Borel-Cantelli Lemma (see, for example, Durrett, 2004), with probability one the event $\{M_{m,c}/m\leq 1-\lambda\}$ only occurs finitely often as $m\to\infty$. That is, $\liminf_m M_{m,c}/m\geq 1-\lambda$ almost surely. Since λ can be arbitrarily close to zero, $M_{m,c}/m\to 1$ almost surely. Observe that this convergence is uniform in $\mathbb{P}$, since T_c only depends on $\mathcal{X}$. ■

Now we proceed to prove the theorem. Given m training samples and m testing samples with $M_{m,c}$ sample pairs, we notice that for these paired samples, both the total testing error and the total testing hinge-loss is upper bounded by

$$\begin{aligned} & \max_{(\delta_1,\cdots,\delta_m)\in\mathcal{N}_0\times\cdots\times\mathcal{N}_0}\sum_{i=1}^{m}\max\big[1-y_i\big(\langle\mathbf{w},\mathbf{x}_i-\delta_i\rangle+b\big),0\big] \\ \leq & cm\|\mathbf{w}\|_2+\sum_{i=1}^{m}\max\big[1-y_i(\langle\mathbf{w},\mathbf{x}_i\rangle+b),0\big], \end{aligned}$$

where $\mathcal{N}_0=\{\delta\,|\,\|\delta\|\leq c\}$. Hence the total classification error of the m testing samples can be upper bounded by

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

and since

$$\max_{\mathbf{x}\in\mathcal{X}}(1-y(\langle\mathbf{w},\mathbf{x}\rangle))\leq\max_{\mathbf{x}\in\mathcal{X}}\Big\{1+|b|+\sqrt{\langle\mathbf{x},\mathbf{x}\rangle\cdot\langle\mathbf{w},\mathbf{w}\rangle}\Big\}=1+|b|+K\|\mathbf{w}\|_2,$$

the accumulated hinge-loss of the total m testing samples is upper bounded by

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

Therefore, the average testing error is upper bounded by

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

and the average hinge loss is upper bounded by

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

Let $\gamma_{m,c} = 1 - M_{m,c}/m$. The proof follows since $M_{m,c}/m \to 1$ almost surely for any $c > 0$. Notice by Inequality (14) we have

$$\mathbb{P}\Big(\gamma_{m,c} \geq \lambda\Big) \leq \exp\Big(-m\lambda^2/8 + (T_c+1)\log 2\Big), \tag{15}$$

that is, the convergence is uniform in $\mathbb{P}$.

We have shown that the average testing error is upper bounded. The final step is to show that this implies that in fact the random variable given by the conditional expectation (conditioned on the training sample) of the error is bounded almost surely as in the statement of the theorem. To make things precise, consider a fixed m, and let $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$ generate the m training samples and m testing samples, respectively, and for shorthand let $\mathcal{T}^m$ denote the random variable of the first m training samples. Let us denote the probability measures for the training by ρ_1 and the testing samples by ρ_2. By independence, the joint measure is given by the product of these two. We rely on this property in what follows. Now fix a λ and a $c > 0$. In our new notation, Equation (15) now reads:

$$\begin{aligned}\int_{\Omega_1}\int_{\Omega_2} \mathbf{1}\big\{\gamma_{m,c}(\omega_1,\omega_2) \geq \lambda\big\}\, d\rho_2(\omega_2)\, d\rho_1(\omega_1) &= \mathbb{P}\Big(\gamma_{m,c}(\omega_1,\omega_2) \geq \lambda\Big) \\ &\leq \exp\Big(-m\lambda^2/8 + (T_c+1)\log 2\Big).\end{aligned}$$

We now bound $\mathbb{P}_{\omega_1}(\mathbb{E}_{\omega_2}[\gamma_{m,c}(\omega_1,\omega_2)\,|\,\mathcal{T}^m] > \lambda)$, and then use Borel-Cantelli to show that this event can happen only finitely often. We have:

$$\begin{aligned}
& \mathbb{P}_{\omega_1}(\mathbb{E}_{\omega_2}[\gamma_{m,c}(\omega_1,\omega_2)\,|\,\mathcal{T}^m] > \lambda) \\
= & \int_{\Omega_1} \mathbf{1}\Big\{\int_{\Omega_2} \gamma_{m,c}(\omega_1,\omega_2)\, d\rho_2(\omega_2) > \lambda\Big\}\, d\rho_1(\omega_1) \\
\leq & \int_{\Omega_1} \mathbf{1}\Big\{\Big[\int_{\Omega_2} \gamma_{m,c}(\omega_1,\omega_2)\mathbf{1}(\gamma_{m,c}(\omega_1,\omega_2) \leq \lambda)\, d\rho_2(\omega_2) + \\
& \int_{\Omega_2} \gamma_{m,c}(\omega_1,\omega_2)\mathbf{1}(\gamma_{m,c}(\omega_1,\omega_2) > \lambda)\, d\rho_2(\omega_2)\Big] \geq 2\lambda\Big\} d\rho_1(\omega_1) \\
\leq & \int_{\Omega_1} \mathbf{1}\Big\{\Big[\int_{\Omega_2} \lambda\mathbf{1}(\lambda(\omega_1,\omega_2) \leq \lambda)\, d\rho_2(\omega_2) + \\
& \int_{\Omega_2} \mathbf{1}(\gamma_{m,c}(\omega_1,\omega_2) > \lambda)\, d\rho_2(\omega_2)\Big] \geq 2\lambda\Big\} d\rho_1(\omega_1) \\
\leq & \int_{\Omega_1} \mathbf{1}\Big\{\Big[\lambda + \int_{\Omega_2} \mathbf{1}(\gamma_{m,c}(\omega_1,\omega_2) > \lambda)\, d\rho_2(\omega_2)\Big] \geq 2\lambda\Big\} d\rho_1(\omega_1) \\
= & \int_{\Omega_1} \mathbf{1}\Big\{\int_{\Omega_2} \mathbf{1}(\gamma_{m,c}(\omega_1,\omega_2) > \lambda)\, d\rho_2(\omega_2) \geq \lambda\Big\} d\rho_1(\omega_1).
\end{aligned}$$

Here, the first equality holds because training and testing samples are independent, and hence the joint measure is the product of ρ_1 and ρ_2. The second inequality holds because $\gamma_{m,c}(\omega_1,\omega_2) \leq 1$ everywhere. Further notice that

$$\begin{aligned}
& \int_{\Omega_1}\int_{\Omega_2} \mathbf{1}\big\{\gamma_{m,c}(\omega_1,\omega_2) \geq \lambda\big\}\, d\rho_2(\omega_2)\, d\rho_1(\omega_1) \\
& \geq \int_{\Omega_1} \lambda\mathbf{1}\Big\{\int_{\Omega_2} \mathbf{1}\big(\gamma_{m,c}(\omega_1,\omega_2) \geq \lambda\big)\, d\rho(\omega_2) > \lambda\Big\}\, d\rho_1(\omega_1).
\end{aligned}$$

Thus we have

$$\mathbb{P}(\mathbb{E}_{\omega_2}(\gamma_{m,c}(\omega_1,\omega_2))>\lambda)\leq \mathbb{P}\Big(\gamma_{m,c}(\omega_1,\omega_2)\geq\lambda\Big)/\lambda\leq \exp\Big(-m\lambda^2/8+(T_c+1)\log 2\Big)/\lambda.$$

For any λ and c, summing up the right hand side over $m=1$ to ∞ is finite, hence the theorem follows from the Borel-Cantelli lemma. ■

Remark 10 We note that M_m/m converges to 1 almost surely even if $\mathcal{X}$ is not bounded. Indeed, to see this, fix $\varepsilon>0$, and let $\mathcal{X}'\subseteq\mathcal{X}$ be a bounded set such that $\mathbb{P}(\mathcal{X}')>1-\varepsilon$. Then, with probability one,

$$\#(\text{unpaired samples in}\mathcal{X}')/m\to 0,$$

by Lemma 9. In addition,

$$\max\big(\#(\text{training samples not in }\mathcal{X}'),\#(\text{testing samples not in }\mathcal{X}')\big)/m\to\varepsilon.$$

Notice that

$$\begin{aligned}M_m\geq m&-\#(\text{unpaired samples in }\mathcal{X}')\\&-\max\big(\#(\text{training samples not in }\mathcal{X}'),\#(\text{testing samples not in }\mathcal{X}')\big).\end{aligned}$$

Hence

$$\lim_{m\to\infty}M_m/m\geq 1-\varepsilon,$$

almost surely. Since ε is arbitrary, we have $M_m/m\to 1$ almost surely.

Next, we prove an analog of Theorem 8 for the kernelized case, and then show that these two imply statistical consistency of linear and kernelized SVMs. Again, 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 $\mathcal{X}\times\{-1,+1\}$.

Theorem 11 *Denote* $K\triangleq\max_{\mathbf{x}\in\mathcal{X}}k(\mathbf{x},\mathbf{x})$. *Suppose there exists* $\rho>0$ *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 $\{\gamma_{m,c}\}$ *such that:*

1. $\forall c>0$, $\lim_{m\to\infty}\gamma_{m,c}=0$ *almost surely, and the convergence is uniform in* $\mathbb{P}$*;*
2. *the following bounds on the Bayes loss and the hinge loss hold uniformly for all* $(\mathbf{w},b)\in\mathcal{H}\times\mathbb{R}$

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

Proof As in the proof of Theorem 8, we generate a set of m testing samples and m training samples, and then lower-bound the number of samples that can form a *sample pair* in the feature-space; that is, a pair consisting of a training sample $(\mathbf{x}, y)$ and a testing sample $(\mathbf{x}', y')$ such that $y = y'$ and $\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|_{\mathcal{H}} \leq c$. In contrast to the finite-dimensional sample space, the feature space may be infinite dimensional, and thus our decomposition may have an infinite number of "bricks." In this case, our multinomial random variable argument used in the proof of Lemma 9 breaks down. 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 $\alpha > 0$. Now notice that by Lemma 7, if a testing sample $\mathbf{x}$ and a training sample $\mathbf{x}'$ belong to a "brick" with length of each side $\min(\rho/\sqrt{n}, f^{-1}(c^2)/\sqrt{n})$ in the *sample space* (see the proof of Lemma 9), $\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|_{\mathcal{H}} \leq c$. 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 brick in the sample space. We can cover $\mathcal{X}$ with finitely many (denoted as T_c) such bricks since $f^{-1}(c^2) > 0$. Then, a similar argument as in Lemma 9 shows that the ratio of samples that form pairs in a brick converges to 1 as m increases. Further notice that for M paired samples, the total testing error and hinge-loss are both upper-bounded by

$$cM\|\mathbf{w}\|_{\mathcal{H}} + \sum_{i=1}^{M} \max\left[1 - y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i)\rangle + b), 0\right].$$

The rest of the proof is identical to Theorem 8. In particular, Inequality (15) still holds. ■

Note that the condition in Theorem 11 is satisfied by most commonly used kernels, for example, homogeneous polynominal kernels and Gaussian radial basis functions. This condition requires that the feature mapping is "smooth" and hence preserves "locality" of the disturbance, that is, small disturbance in the sample space guarantees the corresponding disturbance in the feature space is also small. It is easy to construct non-smooth kernel functions which do not generalize well. For example, consider the following kernel:

$$k(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 & \mathbf{x} = \mathbf{x}'; \\ 0 & \mathbf{x} \neq \mathbf{x}'. \end{cases}$$

A standard RKHS regularized SVM using this kernel leads to a decision function

$$\text{sign}(\sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b),$$

which equals $\text{sign}(b)$ 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 made arbitrarily small by fine-tuning the parameter.

5.1 Convergence to Bayes Risk

Next we relate the results of Theorem 8 and Theorem 11 to the standard consistency notion, that is, convergence to the Bayes Risk (Steinwart, 2005). The key point of interest in our proof is the use of a robustness condition in place of a VC-dimension or stability condition used in Steinwart (2005). The proof in Steinwart (2005) has 4 main steps. They show: (i) there always exists a minimizer to

the expected regularized (kernel) hinge loss; (ii) the expected regularized hinge loss of the minimizer converges to the expected hinge loss as the regularizer goes to zero; (iii) if a sequence of functions asymptotically have optimal expected hinge loss, then they also have optimal expected loss; and (iv) the expected hinge loss of the minimizer of the regularized *training* hinge loss concentrates around the empirical regularized hinge loss. In Steinwart (2005), this final step, (iv), is accomplished using concentration inequalities derived from VC-dimension considerations, and stability considerations.

Instead, we use our robustness-based results of Theorem 8 and Theorem 11 to replace these approaches (Lemmas 3.21 and 3.22 in Steinwart 2005) in proving step (iv), and thus to establish the main result.

Recall that a classifier is a rule that assigns to every training set $T = \{\mathbf{x}_i, y_i\}_{i=1}^m$ a measurable function f_T. The risk of a measurable function $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$\mathcal{R}_{\mathbb{P}}(f) \triangleq \mathbb{P}(\{\mathbf{x}, y : \operatorname{sign} f(\mathbf{x}) \neq y\}).$$

The smallest achievable risk

$$\mathcal{R}_{\mathbb{P}} \triangleq \inf\{\mathcal{R}_{\mathbb{P}}(f) | f : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$$

is called the *Bayes Risk* of $\mathbb{P}$. A classifier is said to be strongly uniformly consistent if for all distributions P on $\mathcal{X} \times [-1, +1]$, the following holds almost surely.

$$\lim_{m\to\infty} \mathcal{R}_{\mathbb{P}}(f_T) = \mathcal{R}_{\mathbb{P}}.$$

Without loss of generality, we only consider the kernel version. Recall a definition from Steinwart (2005).

Definition 12 *Let $C(\mathcal{X})$ be the set of all continuous functions defined on a compact metric space $\mathcal{X}$. Consider the mapping $I : \mathcal{H} \to C(\mathcal{X})$ defined by $I\mathbf{w} \triangleq \langle \mathbf{w}, \Phi(\cdot)\rangle$. If I has a dense image, we call the kernel* universal.

Roughly speaking, if a kernel is universal, then the corresponding RKHS is rich enough to satisfy the condition of step (ii) above.

Theorem 13 *If a kernel satisfies the condition of Theorem 11, and is universal, then the Kernel SVM with $c \downarrow 0$ sufficiently slowly is strongly uniformly consistent.*

Proof We first introduce some notation, largely following Steinwart (2005). For some probability measure μ and $(\mathbf{w}, b) \in \mathcal{H} \times \mathbb{R}$,

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

is the expected hinge-loss under probability μ, and

$$R^c_{L,\mu}((\mathbf{w}, b)) \triangleq c\|\mathbf{w}\|_{\mathcal{H}} + \mathbb{E}_{(\mathbf{x},y)\sim\mu}\big\{\max(0, 1 - y(\langle \mathbf{w}, \Phi(\mathbf{x})\rangle + b))\big\}$$

is the regularized expected hinge-loss. Hence $R_{L,\mathbb{P}}(\cdot)$ and $R^c_{L,\mathbb{P}}(\cdot)$ are the expected hinge-loss and regularized expected hinge-loss under the generating probability $\mathbb{P}$. If μ is the empirical distribution of m samples, we write $R_{L,m}(\cdot)$ and $R^c_{L,m}(\cdot)$ respectively. Notice $R^c_{L,m}(\cdot)$ is the objective function of the SVM. Denote its solution by $f_{m,c}$, that is, the classifier we get by running SVM with m samples

and parameter c. Further denote by $f_{\mathbb{P},c} \in \mathcal{H} \times \mathbb{R}$ the minimizer of $R^c_{L,\mathbb{P}}(\cdot)$. The existence of such a minimizer is proved in Lemma 3.1 of Steinwart (2005) (step (i)). Let

$$\mathcal{R}_{L,\mathbb{P}} \triangleq \min_{f \text{ measurable}} \mathbb{E}_{\mathbf{x},y\sim\mathbb{P}}\Big\{ \max\big(1 - y f(\mathbf{x}), 0\big)\Big\},$$

that is, the smallest achievable hinge-loss for all measurable functions.

The main content of our proof is to use Theorems 8 and 11 to prove step (iv) in Steinwart (2005). In particular, we show: if $c \downarrow 0$ "slowly", we have with probability one

$$\lim_{m\to\infty} R_{L,\mathbb{P}}(f_{m,c}) = \mathcal{R}_{L,\mathbb{P}}. \tag{16}$$

To prove Equation (16), denote by $\mathbf{w}(f)$ and $b(f)$ as the weight part and offset part of any classifier f. Next, we bound the magnitude of $f_{m,c}$ by using $R^c_{L,m}(f_{m,c}) \le R^c_{L,m}(\mathbf{0},0) \le 1$, which leads to

$$\|\mathbf{w}(f_{m,c})\|_{\mathcal{H}} \le 1/c$$

and

$$|b(f_{m,c})| \le 2 + K\|\mathbf{w}(f_{m,c})\|_{\mathcal{H}} \le 2 + K/c.$$

From Theorem 11 (note that the bound holds uniformly for all $(\mathbf{w}, b)$), we have

$$\begin{aligned}
R_{L,\mathbb{P}}(f_{m,c}) &\le \gamma_{m,c}[1 + K\|\mathbf{w}(f_{m,c})\|_{\mathcal{H}} + |b|] + R^c_{L,m}(f_{m,c}) \\
&\le \gamma_{m,c}[3 + 2K/c] + R^c_{L,m}(f_{m,c}) \\
&\le \gamma_{m,c}[3 + 2K/c] + R^c_{L,m}(f_{\mathbb{P},c}) \\
&= \mathcal{R}_{L,\mathbb{P}} + \gamma_{m,c}[3 + 2K/c] + \big\{R^c_{L,m}(f_{\mathbb{P},c}) - R^c_{L,\mathbb{P}}(f_{\mathbb{P},c})\big\} + \big\{R^c_{L,\mathbb{P}}(f_{\mathbb{P},c}) - \mathcal{R}_{L,\mathbb{P}}\big\} \\
&= \mathcal{R}_{L,\mathbb{P}} + \gamma_{m,c}[3 + 2K/c] + \big\{R_{L,m}(f_{\mathbb{P},c}) - R_{L,\mathbb{P}}(f_{\mathbb{P},c})\big\} + \big\{R^c_{L,\mathbb{P}}(f_{\mathbb{P},c}) - \mathcal{R}_{L,\mathbb{P}}\big\}.
\end{aligned}$$

The last inequality holds because $f_{m,c}$ minimizes $R^c_{L,m}$.

It is known (Steinwart, 2005, Proposition 3.2) (step (ii)) that if the kernel used is rich enough, that is, universal, then

$$\lim_{c\to 0} R^c_{L,\mathbb{P}}(f_{\mathbb{P}}, c) = \mathcal{R}_{L,\mathbb{P}}.$$

For fixed $c > 0$, we have

$$\lim_{m\to\infty} R_{L,m}(f_{\mathbb{P},c}) = R_{L,\mathbb{P}}(f_{\mathbb{P},c}),$$

almost surely due to the strong law of large numbers (notice that $f_{\mathbb{P},c}$ is a fixed classifier), and $\gamma_{m,c}[3 + 2K/c] \to 0$ almost surely. Notice that neither convergence rate depends on $\mathbb{P}$. Therefore, if $c \downarrow 0$ sufficiently slowly,[3] we have almost surely

$$\lim_{m\to\infty} R_{L,\mathbb{P}}(f_{m,c}) \le \mathcal{R}_{L,\mathbb{P}}.$$

Now, for any m and c, we have $R_{L,\mathbb{P}}(f_{m,c}) \ge \mathcal{R}_{L,\mathbb{P}}$ by definition. This implies that Equation (16) holds almost surely, thus giving us step (iv).

Finally, Proposition 3.3. of Steinwart (2005) shows step (iii), namely, approximating hinge loss is sufficient to guarantee approximation of the Bayes loss. Thus Equation (16) implies that the risk

3. For example, we can take $\{c(m)\}$ be the smallest number satisfying $c(m) \ge m^{-1/8}$ and $T_{c(m)} \le m^{1/8}/\log 2 - 1$. Inequality (15) thus leads to $\sum_{m=1}^{\infty} P(\gamma_{m,c(m)}/c(m) \ge m^{1/4}) \le +\infty$ which implies uniform convergence of $\gamma_{m,c(m)}/c(m)$.

of function $f_{m,c}$ converges to Bayes risk. ■

Before concluding this section, we remark that although we focus in this paper the hinge-loss function and the RKHS norm regularizer, the robustness approach to establish consistency can be generalized to other regularization schemes and loss functions. Indeed, throughout the proof we only require that the regularized loss (that is, the training loss plus the regularization penalty) is an upper bound of the minimax error with respect to certain set-inclusive uncertainty. This is a property satisfied by many classification algorithms even though an exact equivalence relationship similar to the one presented in this paper may not exist. This suggests using the robustness view to derive sharp sample complexity bounds for a broad class of algorithms (e.g., Steinwart and Christmann, 2008).

6. Concluding Remarks

This work considers the relationship between robust 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. The interpretation of this link is that norm-based regularization essentially builds in a robustness to sample noise whose probability level sets are symmetric unit balls 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.

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. In particular, for "smooth" feature mappings, the robustness to disturbance in the observation space is guaranteed and hence SVMs achieve consistency. On the other-hand, certain "non-smooth" feature mappings fail to be consistent simply because for such kernels the robustness in the feature-space (guaranteed by the regularization process) does not imply robustness in the observation space.

Acknowledgments

We thank the editor and three anonymous reviewers for significantly improving the accessibility of this manuscript. We also benefited from comments from participants in ITA 2008 and at a NIPS 2008 workshop on optimization. This research was partially supported by the Canada Research Chairs Program, by the Israel Science Foundation (contract 890015), by a Horev Fellowship, by NSF Grants EFRI-0735905, CNS-0721532, and a grant from DTRA.

Appendix A.

In this appendix we show that for RBF kernels, it is possible to relate robustness in the feature space and robustness in the sample space more directly.

Theorem 14 *Suppose the Kernel function has the form $k(\mathbf{x},\mathbf{x}') = f(\|\mathbf{x}-\mathbf{x}'\|)$, with $f:\mathbb{R}^+\to\mathbb{R}$ a decreasing function. Denote by $\mathcal{H}$ the RKHS space of $k(\cdot,\cdot)$ and $\Phi(\cdot)$ the corresponding feature mapping. Then we have for any $\mathbf{x}\in\mathbb{R}^n$, $\mathbf{w}\in\mathcal{H}$ and $c>0$,*

$$\sup_{\|\delta\|\le c}\langle \mathbf{w}, \Phi(\mathbf{x}-\delta)\rangle = \sup_{\|\delta_\phi\|_{\mathcal{H}}\le\sqrt{2f(0)-2f(c)}}\langle \mathbf{w}, \Phi(\mathbf{x})+\delta_\phi\rangle.$$

Proof We show that the left-hand-side is not larger than the right-hand-side, and vice versa.

First we show

$$\sup_{\|\delta\|\le c}\langle \mathbf{w}, \Phi(\mathbf{x}-\delta)\rangle \le \sup_{\|\delta_\phi\|_{\mathcal{H}}\le\sqrt{2f(0)-2f(c)}}\langle \mathbf{w}, \Phi(\mathbf{x})-\delta_\phi\rangle. \tag{17}$$

We notice that for any $\|\delta\|\le c$, we have

$$\begin{aligned}
&\langle \mathbf{w}, \Phi(\mathbf{x}-\delta)\rangle\\
=&\Big\langle \mathbf{w}, \Phi(\mathbf{x})+\big(\Phi(\mathbf{x}-\delta)-\Phi(\mathbf{x})\big)\Big\rangle\\
=&\langle \mathbf{w}, \Phi(\mathbf{x})\rangle + \langle \mathbf{w}, \Phi(\mathbf{x}-\delta)-\Phi(\mathbf{x})\rangle\\
\le&\langle \mathbf{w}, \Phi(\mathbf{x})\rangle + \|\mathbf{w}\|_{\mathcal{H}}\cdot\|\Phi(\mathbf{x}-\delta)-\Phi(\mathbf{x})\|_{\mathcal{H}}\\
\le&\langle \mathbf{w}, \Phi(\mathbf{x})\rangle + \|\mathbf{w}\|_{\mathcal{H}}\sqrt{2f(0)-2f(c)}\\
=&\sup_{\|\delta_\phi\|_{\mathcal{H}}\le\sqrt{2f(0)-2f(c)}}\langle \mathbf{w}, \Phi(\mathbf{x})-\delta_\phi\rangle.
\end{aligned}$$

Taking the supremum over δ establishes Inequality (17).

Next, we show the opposite inequality,

$$\sup_{\|\delta\|\le c}\langle \mathbf{w}, \Phi(\mathbf{x}-\delta)\rangle \ge \sup_{\|\delta_\phi\|_{\mathcal{H}}\le\sqrt{2f(0)-2f(c)}}\langle \mathbf{w}, \Phi(\mathbf{x})-\delta_\phi\rangle. \tag{18}$$

If $f(c)=f(0)$, then Inequality 18 holds trivially, hence we only consider the case that $f(c)<f(0)$. Notice that the inner product is a continuous function in $\mathcal{H}$, hence for any $\varepsilon>0$, there exists a δ'_ϕ such that

$$\langle \mathbf{w}, \Phi(\mathbf{x})-\delta'_\phi\rangle > \sup_{\|\delta_\phi\|_{\mathcal{H}}\le\sqrt{2f(0)-2f(c)}}\langle \mathbf{w}, \Phi(\mathbf{x})-\delta_\phi\rangle-\varepsilon;\quad \|\delta'_\phi\|_{\mathcal{H}}<\sqrt{2f(0)-2f(c)}.$$

Recall that the RKHS space is the completion of the feature mapping, thus there exists a sequence of $\{\mathbf{x}'_i\}\in\mathbb{R}^n$ such that

$$\Phi(\mathbf{x}'_i)\to\Phi(\mathbf{x})-\delta'_\phi, \tag{19}$$

which is equivalent to

$$\big(\Phi(\mathbf{x}'_i)-\Phi(\mathbf{x})\big)\to-\delta'_\phi.$$

This leads to

$$\begin{aligned}
&\lim_{i\to\infty}\sqrt{2f(0)-2f(\|\mathbf{x}'_i-\mathbf{x}\|)}\\
=&\lim_{i\to\infty}\|\Phi(\mathbf{x}'_i)-\Phi(\mathbf{x})\|_{\mathcal{H}}\\
=&\|\delta'_\phi\|_{\mathcal{H}}<\sqrt{2f(0)-2f(c)}.
\end{aligned}$$

Since f is decreasing, we conclude that $\|\mathbf{x}_i' - \mathbf{x}\| \leq c$ holds except for a finite number of i. By (19) we have

$$\langle \mathbf{w}, \Phi(\mathbf{x}_i') \rangle \to \langle \mathbf{w}, \Phi(\mathbf{x}) - \delta_\phi' \rangle > \sup_{\|\delta_\phi\|_{\mathcal{H}} \leq \sqrt{2f(0)-2f(c)}} \langle \mathbf{w}, \Phi(\mathbf{x}) - \delta_\phi \rangle - \varepsilon,$$

which means

$$\sup_{\|\delta\| \leq c} \langle \mathbf{w}, \Phi(\mathbf{x} - \delta) \rangle \geq \sup_{\|\delta_\phi\|_{\mathcal{H}} \leq \sqrt{2f(0)-2f(c)}} \langle \mathbf{w}, \Phi(\mathbf{x}) - \delta_\phi \rangle - \varepsilon.$$

Since ε is arbitrary, we establish Inequality (18).

Combining Inequality (17) and Inequality (18) proves the theorem. ■

References

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

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

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

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

D. Bertsimas and A. Fertis. Personal Correspondence, March 2008.

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

C. Bhattacharyya. Robust classification of noisy data using second order cone programming approach. In *Proceedings International Conference on Intelligent Sensing and Information Processing*, pages 433–438, Chennai, India, 2004.

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

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

J. Bi and T. Zhang. Support vector classification with input data uncertainty. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems (NIPS17)*, Cambridge, MA, 2004. MIT Press.

C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108. URL http://www.mitpressjournals.org/doi/abs/10.1162/neco.1995.7.1.108.

B. E. Boser, I. M. Guyon, and V. N. 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.

O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

A. Christmann and I. Steinwart. On robustness properties of convex risk minimization methods for pattern recognition. *The Journal of Machine Learning Research*, 5:1007–1034, 2004.

A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression in convex risk minimization. *Bernoulli*, 13(3):799–819, 2007.

A. Christmann and A. Van Messem. Bouligand derivatives and robustness of support vector machines for regression. *The Journal of Machine Learning Research*, 9:915–936, 2008.

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

R. Durrett. *Probability: Theory and Examples*. Duxbury Press, 2004.

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.

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

A. Globerson and S. Roweis. Nightmare at test time: Robust learning by feature deletion. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*, pages 353–360, New York, NY, USA, 2006. ACM Press.

F. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York, 1986.

P. J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

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

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

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *UAI-2002: Uncertainty in Artificial Intelligence*, pages 275–282, 2002.

G. R. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *The Journal of Machine Learning Research*, 3:555–582, 2003.

R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, New York, 2006.

S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, 2006.

T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, 1987.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

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

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

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

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

C. H. Teo, A. Globerson, S. Roweis, and A. J. Smola. Convex learning with invariances. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1489–1496, Cambridge, MA, 2008. MIT Press.

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

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 2000.

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

V. N. 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.

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

H. Xu, C. Caramanis, and S. Mannor. Robust regression and Lasso. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1801–1808, 2009.