A Simple Geometric Interpretation of SVM using Stochastic Adversaries

Roi Livni
ELSC-ICNC Edmond & Lily Safra
Center for Brain Sciences
The Hebrew University
Jerusalem, Israel

Koby Crammer
Department of
Electrical Engineering
The Technion
Haifa, Israel

Amir Globerson
School of Computer Science
and Engineering
The Hebrew University
Jerusalem, Israel

Abstract

We present a minimax framework for classification that considers stochastic adversarial perturbations to the training data. We show that for binary classification it is equivalent to SVM, but with a very natural interpretation of regularization parameter. In the multiclass case, we obtain that our formulation is equivalent to regularizing the hinge loss with the maximum norm of the weight vector (i.e., the two-infinity norm). We test this new regularization scheme and show that it is competitive with the Frobenius regularization commonly used for multiclass SVM. We proceed to analyze various forms of stochastic perturbations and obtain compact optimization problems for the optimal classifiers. Taken together, our results illustrate the advantage of using stochastic perturbations rather than deterministic ones, as well as offer a simple geometric interpretation for SVM optimization.

1 Introduction

One of the most common approaches to classification is to minimize the regularized empirical hinge loss [22]. The most common regularization is the squared ℓ_2 norm. One of the motivations for this approach comes from the linearly separable case, in which minimizing the norm subject to classifying the points correctly is equivalent to maximizing the separation margin. However, minimizing the ℓ_2 regularized hinge loss no longer has this nice geometric interpretation.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

In this work, we provide a simple geometric view of regularized hinge loss minimization, as well as other regularized loss optimization scenarios. We show that this geometric interpretation arises naturally when casting learning in robust optimization terms, and specifically in robustness to stochastic perturbations.

In the typical robust setting, the learning algorithm seeks a classifier that will perform well not only on the training examples, but also on their perturbed versions. The optimization follows a minimax setting where an adversary has access to the classifier $\boldsymbol{w}$ and can modify the input points $\boldsymbol{x}$ to maximize the loss incurred by the classifier. In all previous works in this setting [1, 12, 13, 24, 10, 6, 8]) the adversary has the power to move the samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to a set $\mathcal{S}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ defined in advance (e.g., the set of balls of radius R around each sample point). The goal of the learning algorithm is then to minimize the worst case loss.

Here we consider a stochastic variant of this setting, which turns out to be much more closely related to standard regularized loss minimization. We employ a stochastic adversary which uses a distribution over space to sample new perturbed points. On the one hand, we maintain the adversarial nature of the formulation, but on the other hand a stochastic adversary is conceptually weaker than a deterministic one, since it cannot choose the worst point, but rather needs to spread points stochastically. Clearly, the distribution chosen by the adversary needs to be limited somehow, and we consider several natural choices of distribution sets, focusing on those bounding the expected deviation from the original point $\boldsymbol{x}$.

As in standard robust classification, we consider a minimax setting where the learner first chooses a weight vector, and the adversary reacts to it by choosing a worst case distribution over inputs. The goal of the learner is to minimize the maximum expected hinge loss, where expectation is taken over the adversarial

distribution. In general, solving the minimax problem with stochastic adversaries is quite challenging since the adversary gets to optimize over the set of all probability measures. We show how to overcome this difficulty and obtain several elegant equivalent formulations of the problem.

Our first result is to show that for binary classification the optimal strategy is equivalent to learning via SVMs. Interestingly, for the multiclass case, our formulation yields an alternative regularization to current existing versions of multi-class SVMs. Instead of using the standard $\ell_{2,2}$ over the weight vectors (e.g. [27, 5]), our algorithms result in an $\ell_{2,\infty}$ regularization. Our formulation also suggests a natural choice of regularization parameter which often works well in practice. We show empirically that our resulting multiclass method is competitive with SVM classifiers.

Finally, we generalize our analysis to other perturbation models and loss functions, including log-loss and ℓ_1 regularization. Taken together, our results illustrate the utility of using stochastic adversaries, and the simple geometric interpretations that result from this analysis.

2 Problem Formulation

We assume a given set of n labeled examples $\{\boldsymbol{x}_i, y_i\}$ with $i = 1, \ldots, n$. Each example is composed of an input vector $\boldsymbol{x}_i \in \mathbb{R}^d$ and a label from a finite set of size L, $y_i \in \{1, \ldots, L\}$. In this work we consider linear models parameterized by a set of L weight vectors $\boldsymbol{w}_y \in \mathbb{R}^d$ - one vector per possible class, for $y = 1, \ldots, L$ [5]. Given a new input, the predicted label is the one which maximizes the inner-product of the input $\boldsymbol{x}$ and the corresponding weight-vector, $\hat{y} = \arg\max_y \boldsymbol{w}_y \cdot \boldsymbol{x}$.

A common strategy for learning the weights $\boldsymbol{w}_y$ is to minimize an upper bound on the average zero-one loss of the training set. Here we follow the common strategy of bounding the zero-one loss with the hinge loss. We use the following multiclass hinge loss [5]:

$$\ell(\boldsymbol{x}, y; \boldsymbol{w}) = \max_{\bar{y}} \left[\boldsymbol{w}_{\bar{y}} \cdot \boldsymbol{x} - \boldsymbol{w}_y \cdot \boldsymbol{x} + e_{y,\bar{y}}\right] \qquad (1)$$

where $e_{y,\bar{y}}$ is zero if $y = \bar{y}$ and one otherwise.

The standard machine learning approach once a loss is defined, is to minimize a sum of the average loss and a regularization term. The goal of the latter is to prevent over-fitting by biasing the algorithm towards "simple" models. For example, the following multiclass SVM formulation [5] minimizes the sum of the average hinge loss and the squared ℓ_2 norm of the weights,

$$\min_{\boldsymbol{w}} \tfrac{1}{n} \textstyle\sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) + \frac{C}{2} \sum_y \|\boldsymbol{w}_y\|_2^2 \qquad (2)$$

where C is a regularization parameter, used to trade-off simplicity and accuracy.

In our formulation we do not introduce such a regularization. Rather, we employ a minimax formulation and we shall see how regularization naturally follows from it.

Our formulation requires that not only should the learned classifier perform well on the training data, but it should also be robust with respect to perturbations of the input points $\boldsymbol{x}_i$. Unlike previous approaches for robust classification [12, 13, 10] we focus on perturbations that are stochastic in nature and not deterministic. We model perturbations of an input $\boldsymbol{x}$ via a conditional distribution $p(\bar{\boldsymbol{x}}|\boldsymbol{x})$, and formalize the learning process as a minimization of the *expected loss* w.r.t. this distribution. For this setup to be well-defined the distribution $p(\bar{\boldsymbol{x}}|\boldsymbol{x})$ is constrained to belong to a subset of the possible measures on $\mathbb{R}^d$.

Below we sketch two alternative definitions of the set of distributions. In both cases we enforce the expectation of $\bar{\boldsymbol{x}}$ with respect to $p(\bar{\boldsymbol{x}}|\boldsymbol{x})$ to be $\boldsymbol{x}$ (i.e., the original input point). Additionally, we limit the *spread* of $p(\bar{\boldsymbol{x}}|\boldsymbol{x})$ so that it represents perturbations that are local with high-probability. We formalize this limit via a bound on the average distance of $\bar{\boldsymbol{x}}$ from $\boldsymbol{x}$, and we consider either the Euclidean distance or its square. Since the square function and the expectation operator are not commutative, these two alternatives are not equivalent in general. In summary, we consider perturbation sets of the type

$$\mathcal{S}(\boldsymbol{x}; \sigma, f) = \left\{ p(\bar{\boldsymbol{x}}|\boldsymbol{x}) \in \mathcal{P} : \begin{array}{l} E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}\left[\bar{\boldsymbol{x}}\right] = \boldsymbol{x} \\ E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}\left[f(\bar{\boldsymbol{x}}, \boldsymbol{x})\right] = \sigma \end{array} \right\}$$

where $\mathcal{P}$ is the set of all Borel probability measures over $\mathbb{R}^d$. The two cases we address are $\mathcal{S}_{\ell_2}$ which corresponds to $\mathcal{S}(\boldsymbol{x}; \sigma, f)$ with $f(\bar{\boldsymbol{x}}, \boldsymbol{x}) = \|\bar{\boldsymbol{x}} - \boldsymbol{x}\|_2$, and $\mathcal{S}_{\ell_2^2}$ which corresponds to $\mathcal{S}(\boldsymbol{x}; \sigma, f)$ with $f(\bar{\boldsymbol{x}}, \boldsymbol{x}) = \|\bar{\boldsymbol{x}} - \boldsymbol{x}\|_2^2$.

These sets induce two corresponding stochastically-robust optimization problems:

$$\min_{\boldsymbol{w}} \tfrac{1}{n} \textstyle\sum_i \max_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i) \in \mathcal{S}(\boldsymbol{x}_i; \sigma, f)} E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i)}\left[\ell(\bar{\boldsymbol{x}}, y_i; \boldsymbol{w})\right] \qquad (3)$$

When $\mathcal{S}_{\ell_2}$ is considered, we denote the problem by $\mathrm{RSVM}_2(\sigma)$, and when $\mathcal{S}_{\ell_2^2}$ is considered, we denote the problem by $\mathrm{RSVM}_2^2(\sigma)$.

3 Solving the classification problems

Optimization problems of the form in Eq. 3 are not trivial, as the maximization part is over an infinite dimensional space (i.e., the space of Borel measures on $\mathbb{R}^d$). In what follows, we obtain compact tractable

forms for solving the robust classification problems in the previous section by working with their convex dual problems. We make use of the observation that the maximization problems correspond to linear programs over variables $p(\bar{\boldsymbol{x}}|\boldsymbol{x})$ (e.g., see [15, 26]). Specifically, they are of the form:

$$\begin{array}{ll} \max\limits_{p\in\mathcal{P}} & E_p[f_0(\bar{\boldsymbol{x}})] \\ s.t. & E_p[f_j(\bar{\boldsymbol{x}})] = a_j \quad j = 1,\ldots,m. \end{array} \tag{4}$$

where $f_j(\bar{\boldsymbol{x}})$ are functions on $\bar{\boldsymbol{x}}$. For example, in $\text{RSVM}_2(\sigma)$ the i^{th} point problem corresponds to:

$$f_0(\bar{\boldsymbol{x}}, y_i; \boldsymbol{w}) = \ell(\bar{\boldsymbol{x}}, y_i; \boldsymbol{w}), f_1(\bar{\boldsymbol{x}}) = \bar{\boldsymbol{x}}, f_2(\bar{\boldsymbol{x}}) = \|\bar{\boldsymbol{x}} - \boldsymbol{x}_i\|_2$$

and $a_1 = \boldsymbol{x}_i, a_2 = \sigma$. The dual problem is a linear optimization problem with infinitely many constraints, and variables $z_0, \ldots, z_m$:

$$\begin{array}{ll} \min. & z_0 + \sum_{j=1}^m a_j z_j \\ \text{s.t.} & z_0 + \sum_{j=1}^m z_j f_j(\bar{\boldsymbol{x}}) \geq f_0(\bar{\boldsymbol{x}}) \quad \forall \bar{\boldsymbol{x}}. \end{array} \tag{5}$$

The dual objective, Eq. 5, is always an upper bound on the primal problem. Moreover, it is known [15] that under mild conditions (which are satisfied by all the problems we consider), the dual problem is a strict upper bound. Specifically, the supremum of the primal problem coincides with the minimum of the dual. As we will see below, analyzing the dual is simpler since the number of constraints can be reduced significantly, to a polynomial size.

3.1 Solving RSVM_2

We now develop RSVM_2 into a more tractable form. The key consequence of our analysis is that this problem is equivalent to SVM in binary classification problems. Furthermore, in the multi class cases it is equivalent to an alternative to standard multi-class SVM.

To present our result, we define the following *new* multiclass optimization problem which will be shown to be equivalent to the minimax problem RSVM_2 (see Theorem 3.1).

$$\underline{(\ell_2)_\infty SVM(\sigma)}: \\ \min_{\boldsymbol{w}} \frac{1}{n}\sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) + \sigma \max_y \|\boldsymbol{w}_y\|_2$$

The objective of $(\ell_2)_\infty SVM(\sigma)$ is a sum of two terms. The first is the average hinge loss on the input examples. The second term is an $\ell_{2,\infty}$ norm on the matrix of weights, this is the maximal value of the Euclidean norm over all L weights vectors, one per label.

Before proving that $(\ell_2)_\infty SVM(\sigma)$ and RSVM_2 are equivalent we mention some implications of this result. First, we claim that in the binary class case, $(\ell_2)_\infty SVM(\sigma)$ is equivalent to the classic ℓ_2 regularized hinge loss. To see this, note we can assume *wlog* that the optimal solution satisfies $\boldsymbol{w}_1 = -\boldsymbol{w}_2$ (since the objective in $\boldsymbol{w}$ is translation invariant so we can shift $\boldsymbol{w}_1, \boldsymbol{w}_2$ so that their sum is zero), and thus $\max_y \|\boldsymbol{w}_y\|_2 = \|\boldsymbol{w}_1\|_2$. So the $\ell_{2,\infty}$ regularization is just ℓ_2 regularization. Note that in SVM the squared ℓ_2 norm is typically used. However, the squared and non-squared cases are equivalent up to the precise value of the regularization constant σ.[1] Thus, we obtain a novel and simple interpretation of binary SVM with ℓ_2 regularization: it is simply the best minimax strategy if one wants to be robust to stochastic perturbation of sample points.

In the multi-class case, we note that $(\ell_2)_\infty SVM(\sigma)$ uses a regularization that is different from the Frobenius norm often used in this context (e.g., see [5]). In the Frobenius case, the regularization is the sum of norms of $\boldsymbol{w}_y$ whereas in $(\ell_2)_\infty SVM(\sigma)$ it is the maximum over these norms. We proceed to state and prove the equivalence between $(\ell_2)_\infty SVM(\sigma)$ and RSVM_2.

Theorem 3.1. *The optimization problem $RSVM_2(\sigma)$ is equivalent to $(\ell_2)_\infty SVM(\sigma)$. Namely, they have the same optimal value and optimal assignment to the parameters $\boldsymbol{w}_y$.*

Proof. Consider the adversary maximization problem for the labeled pair $\boldsymbol{x}, y$ (we begin by considering a sample point, and later consider the whole sample),

$$\begin{array}{ll} \max\limits_{p\in\mathcal{P}} & E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}\left[\ell(\bar{\boldsymbol{x}}; y; \boldsymbol{w})\right] \\ s.t. & E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}\left[\bar{\boldsymbol{x}}\right] = \boldsymbol{x} \\ & E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}\left[\|\bar{\boldsymbol{x}} - \boldsymbol{x}\|_2\right] = \sigma. \end{array} \tag{6}$$

As discussed earlier, the dual problem is given by:

$$\begin{array}{ll} \min\limits_{\alpha,\boldsymbol{\beta},\gamma} & \alpha\sigma + \boldsymbol{\beta}^T\boldsymbol{x} + \gamma \\ \text{s.t.} & \alpha\|\bar{\boldsymbol{x}} - \boldsymbol{x}\|_2 + \boldsymbol{\beta}^T\bar{\boldsymbol{x}} + \gamma \geq \ell(\bar{\boldsymbol{x}}; y; \boldsymbol{w}) \quad \forall \bar{\boldsymbol{x}} \end{array}$$

where $\alpha, \gamma \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^d$. We decompose the hinge-loss to linear objective and constraints, change variables ($\bar{\boldsymbol{x}} \leftarrow \bar{\boldsymbol{x}} - \boldsymbol{x}$) and obtain an equivalent form:

$$\begin{array}{ll} \min\limits_{\alpha,\beta,\gamma} & \alpha\sigma + \boldsymbol{\beta}^T\boldsymbol{x} + \gamma \\ \text{s.t.} & \\ \forall \bar{\boldsymbol{x}}, \bar{y} & \alpha\|\bar{\boldsymbol{x}}\|_2 + \boldsymbol{\beta}^T(\bar{\boldsymbol{x}} + \boldsymbol{x}) + \gamma \geq e_{y,\bar{y}} + \Delta\boldsymbol{w}_{\bar{y}}^T(\bar{\boldsymbol{x}} + \boldsymbol{x}) \end{array} \tag{7}$$

where $\Delta\boldsymbol{w}_{\bar{y}} = \boldsymbol{w}_{\bar{y}} - \boldsymbol{w}_y$. For every $\bar{y}$, two necessary and sufficient conditions for the last constraint to hold are:

$$\boldsymbol{\beta}^T\boldsymbol{x} + \gamma \geq e_{y,\bar{y}} + \Delta\boldsymbol{w}_{\bar{y}}^T\boldsymbol{x} \tag{8}$$

[1] Specifically, since a constraint on ℓ_2 norm can be expressed equivalently as a constraint on ℓ_2^2. For the two formulations to be equivalent different values of σ should be used.

and:

$$\forall \bar{\boldsymbol{x}} \quad \alpha\|\bar{\boldsymbol{x}}\|_2 + (\boldsymbol{\beta} - \Delta\boldsymbol{w}_{\bar{y}})^T\bar{\boldsymbol{x}} \geq 0 \ . \tag{9}$$

Sufficiency follows from summing the constraints in Eq. 9 and Eq. 8. The first constraint is necessary by setting $\bar{\boldsymbol{x}} = 0$. The second constraint is necessary as shown by the following negation argument. If indeed there is some $\bar{\boldsymbol{x}}$ for which Eq. 9 is violated, it will also be violated with $r\bar{\boldsymbol{x}}$ for sufficiently large r such that,

$$r\left(\alpha\|\bar{\boldsymbol{x}}\|_2 + (\boldsymbol{\beta} - \Delta\boldsymbol{w}_{\bar{y}})^T\bar{\boldsymbol{x}}\right) < -\boldsymbol{\beta}^T\boldsymbol{x} - \gamma + e_{y,\bar{y}} + \Delta\boldsymbol{w}_{\bar{y}}^T\boldsymbol{x}$$

which yields

$$\alpha\|r\bar{\boldsymbol{x}}\|_2 + \boldsymbol{\beta}^T r\bar{\boldsymbol{x}} + \boldsymbol{\beta}^T\boldsymbol{x} + \gamma < e_{y,\bar{y}} + \Delta\boldsymbol{w}_{\bar{y}}^T r\bar{\boldsymbol{x}} + \Delta\boldsymbol{w}_{\bar{y}}^T\boldsymbol{x}$$

implying that Eq. 7 is wrong.

Eq. 9 can be replaced with the single constraint:

$$\|\boldsymbol{\beta} - \Delta\boldsymbol{w}_{\bar{y}}\|_2 \leq \alpha \ . \tag{10}$$

To see this, note that substituting $\bar{\boldsymbol{x}} = \Delta\boldsymbol{w}_{\bar{y}} - \boldsymbol{\beta}$ in Eq. 9 yields Eq. 10. On the other hand, Eq. 9 follows from Eq. 10 via the Cauchy-Schwartz inequality. The equivalence can alternatively be derived from the fact that Eq. 9 is equivalent to upper bounding the dual norm of the ℓ_2 norm[2] of $\Delta\boldsymbol{w}_{\bar{y}} - \boldsymbol{\beta}$ by α. Since ℓ_2 is self dual, the equivalence follows.

Plugging Eq. (8,9,10) in Eq. 7 we get,

$$\begin{array}{ll} \min\limits_{\alpha,\boldsymbol{\beta},\gamma} & \alpha\sigma + \boldsymbol{\beta}^T\boldsymbol{x} + \gamma \\ \text{s.t.} & \|\boldsymbol{\beta} - \Delta\boldsymbol{w}_{\bar{y}}\|_2 \leq \alpha \quad \forall\bar{y} \\ \forall\bar{y} & \boldsymbol{\beta}^T\boldsymbol{x} + \gamma \geq e_{y,\bar{y}} + \Delta\boldsymbol{w}_{\bar{y}}^T\boldsymbol{x} \end{array} \tag{11}$$

Optimizing over γ we get:

$$\min_{\alpha,\boldsymbol{\beta}} \alpha\sigma + \ell(\boldsymbol{x}; y; \boldsymbol{w}) \ \text{ s.t. } \ \|\boldsymbol{\beta} - \Delta\boldsymbol{w}_{\bar{y}}\|_2 \leq \alpha \quad \forall\bar{y} \tag{12}$$

Finally, we note that both the objective and the constraint are invariant to adding a constant vector to $\boldsymbol{w}_y$. We thus replace $\boldsymbol{w}_{\bar{y}} \leftarrow \boldsymbol{w}_{\bar{y}} - \boldsymbol{w}_y + \beta$ and get,

$$\min_{\alpha} \alpha\sigma + \ell(\boldsymbol{x}; y; \boldsymbol{w}) \ \text{ s.t. } \ \|\boldsymbol{w}_{\bar{y}}\|_2 \leq \alpha \quad \forall\bar{y}$$

Optimizing over α we get that problem $\text{RSVM}_2(\sigma)$ is equivalent to:

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) + \sigma \max_y \|\boldsymbol{w}_y\|_2 \tag{13}$$

which is of the desired form. □

In general there is no probability distribution that attains the supremum value. In the supplementary file we give a construction of a sequence of probabilities that attain the supremum.

[2]Given a norm $\|\cdot\|$ on a linear space X, recall that the dual norm $\|\cdot\|_*$ on X^* is defined as $\|x^*\|_* = \sup_x \frac{\langle x, x^*\rangle}{\|x\|}$.

3.2 Extension to general norms

Our results until now are only for the special case of a perturbation whose expected value is bounded in ℓ_2 norm. We claim that this result can be generalized to any norm constraint on the perturbation. Specifically, given any norm $\|\cdot\|$, define the corresponding set $\mathcal{S}_{\|\cdot\|}(\boldsymbol{x}; \sigma)$,

$$\mathcal{S}_{\|\cdot\|}(\boldsymbol{x}; \sigma) = \left\{ p(\bar{\boldsymbol{x}}|\boldsymbol{x}) \in \mathcal{P} : \begin{array}{l} E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}\left[\bar{\boldsymbol{x}}\right] = \boldsymbol{x} \\ E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x})}\left[\|\bar{\boldsymbol{x}} - \boldsymbol{x}\|\right] = \sigma \end{array} \right\}$$

In the following theorem, we generalize the case of ℓ_2 constraints to general norms.

Theorem 3.2. *The optimization problem:*

$$\min_{\boldsymbol{w}} \tfrac{1}{n}\textstyle\sum_i \displaystyle\max_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i)\in\mathcal{S}_{\|\cdot\|}(\boldsymbol{x}_i;\sigma)} E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i)}\ell(\bar{\boldsymbol{x}}, y_i; \boldsymbol{w}) \tag{14}$$

is equivalent to: $\min\limits_{\boldsymbol{w}} \frac{1}{n}\sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}) + \sigma \max\limits_y \|\boldsymbol{w}_y\|_*$, *where* $\|\cdot\|_*$ *is the dual norm of* $\|\cdot\|$.

The proof follows the same line as of the previous theorem which states the equivalence between Eq. 9 and Eq. 10. We only need to replace the Euclidean norm (and its dual, the Euclidean norm again) with the specific norm and its dual.

Two interesting cases of the above are: ℓ_1 constraints which translate to $\max_y \|\boldsymbol{w}_y\|_\infty$ regularization, and ℓ_∞ constraints which translate to $\max_y \|\boldsymbol{w}_y\|_1$ regularization [21]. Note that the squared Euclidean norm is not a norm, and thus needs to be analyzed separately (see Sec. 3.4).

3.3 Extension to General Losses

Thus far we only considered the hinge loss as the loss we wish to minimize. Here we generalize our results to a wider set of losses. To obtain results similar to what we had earlier, we consider non-negative convex loss functions ℓ (this is satisfied for almost all typical surrogate losses) with the following two properties:

1. ℓ is translation invariant with respect to $\boldsymbol{w}$ (the matrix with $\boldsymbol{w}_y$ in each row), i.e. if $\boldsymbol{w}' - \boldsymbol{w}$ is a constant row matrix, then $\ell(\boldsymbol{x}, y, \boldsymbol{w}') = \ell(\boldsymbol{x}, y, \boldsymbol{w})$.
2. The second property describes the set of subgradients of the loss function. We require that a vector $\mathbf{r}$ is a subgradient of $\ell(\boldsymbol{x}, y, \boldsymbol{w})$ with respect to $\boldsymbol{x}$ if and only if it can be written as $a\boldsymbol{w}_y + \boldsymbol{v}$ where $\boldsymbol{v}$ is a vector in a set $V(\boldsymbol{w})$, dependent only on $\boldsymbol{w}$ and $a \neq 0$ is a constant scalar.

The following theorem gives a simple characterization of our minimax problem for losses that satisfy the above requirements.

Theorem 3.3. *The optimization problem:*

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \max_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i)\in\mathcal{S}_{\|\cdot\|}(\boldsymbol{x}_i;\sigma)} E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i)}\ell(\bar{\boldsymbol{x}},y_i;\boldsymbol{w}) \quad (15)$$

is equivalent to: $\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \ell(\boldsymbol{x}_i,y_i,\boldsymbol{w}) + \sigma \sup_{V(\boldsymbol{w})} \|v\|_*$, *where* $\|\cdot\|_*$ *is the dual norm of* $\|\cdot\|$.

The theorem has exactly the same form as what we observed for the hinge loss (see Theorem 3.2). Namely, the minimax problem is equivalent to regularized loss minimization. The functional form of the regularization depends on the loss through V.

Proof. First note that since ℓ is translation invariant its subgradients are also translation invariant. Hence if $\boldsymbol{v}\in V(\boldsymbol{w})$ we have, $\boldsymbol{v}-a\boldsymbol{b}\in V(\boldsymbol{w}+\boldsymbol{b})$, for every $\boldsymbol{b}$. As before, considering the dual problem we obtain:

$$\begin{array}{ll}\min\limits_{\alpha,\boldsymbol{\beta},\gamma} & \alpha\sigma+\boldsymbol{\beta}^T\boldsymbol{x}+\gamma \\ \text{s.t.} & \alpha\|\bar{\boldsymbol{x}}-\boldsymbol{x}\|_2+\boldsymbol{\beta}^T\bar{\boldsymbol{x}}+\gamma\geq\ell(\bar{\boldsymbol{x}};y;\boldsymbol{w}) \quad \forall\bar{\boldsymbol{x}}\end{array} \quad (16)$$

ℓ is a convex function in $\boldsymbol{x}$, and demanding that a function f is larger than ℓ at every point is equivalent to demanding that f is larger than every affine function $c+\boldsymbol{v}^T\boldsymbol{x}$, that supports ℓ at some point. Let $\mathcal{A}$ be the set of affine functions that supports $\ell(\cdot,y,\boldsymbol{w})$, and recall that $l(x)\in\mathcal{A}$ means:[3]

$$\begin{aligned} l(\boldsymbol{x}) &= \ell(\boldsymbol{z})+\nabla\ell(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{z}) \\ &= C(\boldsymbol{z})+\nabla\ell(\boldsymbol{z})^T\boldsymbol{x}\end{aligned}$$

where $C(\boldsymbol{z})=\ell(\boldsymbol{z})-\nabla\ell(\boldsymbol{z})^T\boldsymbol{z}$. We now replace the condition in Eq. 16 with the equivalent condition that the RHS is greater than all affine supports of ℓ:[4]

$$\begin{array}{ll}\min\limits_{\alpha,\boldsymbol{\beta},\gamma} & \alpha\sigma+\boldsymbol{\beta}^T\boldsymbol{x}+\gamma \\ \text{s.t.} & \alpha\|\bar{\boldsymbol{x}}\|_2+\beta^T(\bar{\boldsymbol{x}}+\boldsymbol{x})+\gamma\geq C(\boldsymbol{z})+\nabla\ell(\boldsymbol{z})^T(\bar{\boldsymbol{x}}+\boldsymbol{x})\end{array}$$

where the above constraints are $\forall\bar{\boldsymbol{x}},\boldsymbol{z}$. As before, for each subgradient, indexed by $\boldsymbol{z}$, two necessary and sufficient conditions for the corresponding constraints to hold are:

$$\beta^T\boldsymbol{x}+\gamma\geq C(\boldsymbol{z})+\nabla\ell(\boldsymbol{z})^T\boldsymbol{x} \quad (17)$$

and:

$$\forall\bar{\boldsymbol{x}} \quad \alpha\|\bar{\boldsymbol{x}}\|+(\beta-\nabla\ell(\boldsymbol{z}))^T\bar{\boldsymbol{x}}\geq 0\ . \quad (18)$$

Eq. 18 is equivalent to $\alpha\geq\|\beta-\nabla\ell(\boldsymbol{z})\|_*$.

This leads to the problem:

$$\begin{array}{ll}\min\limits_{\alpha,\boldsymbol{\beta},\gamma} & \alpha\sigma+\boldsymbol{\beta}^T\boldsymbol{x}+\gamma \\ \text{s.t.} & \boldsymbol{\beta}^T\boldsymbol{x}+\gamma\geq C(\boldsymbol{z})+\nabla\ell(\boldsymbol{z})^T\boldsymbol{x} \quad \forall\boldsymbol{z} \\ & \alpha\geq\|\beta-\nabla\ell(\boldsymbol{z})\|_* \quad \forall\boldsymbol{z}\end{array}$$

[3]We drop the dependence of ℓ on $y,\boldsymbol{w}$ in what follows for notational convenience.

[4]As before we change variables $\bar{\boldsymbol{x}}-\boldsymbol{x}\to\bar{\boldsymbol{x}}$.

The first constraint says that $\boldsymbol{\beta}^T\boldsymbol{x}+\gamma$ is greater than all affine supports of the loss and hence is equivalent to requiring: $\beta^T\boldsymbol{x}+\gamma\geq\ell(\boldsymbol{x},y,\boldsymbol{w})$. Now, using our assumption of subgradients form we obtain:

$$\begin{array}{ll}\min\limits_{\alpha,\boldsymbol{\beta},\gamma} & \alpha\sigma+\boldsymbol{\beta}^T\boldsymbol{x}+\gamma \\ \text{s.t.} & \boldsymbol{\beta}^T\boldsymbol{x}+\gamma\geq\ell(\boldsymbol{x},y,\boldsymbol{w}) \\ & \alpha\geq\|\beta-a\boldsymbol{w}_y+v\|_* \quad \forall v\in V(\boldsymbol{w})\end{array}$$

which is equivalent to

$$\begin{array}{ll}\min\limits_{\alpha,\boldsymbol{\beta},\gamma} & \alpha\sigma+\ell(\boldsymbol{x},y,\boldsymbol{w}) \\ \text{s.t.} & \alpha\geq\|\beta-a\boldsymbol{w}_y+v\|_* \quad \forall v\in V(\boldsymbol{w})\ .\end{array}$$

The desired result follows from a change of variables as in Sec. 3.1. □

To see how this result applies to the hinge loss, note that the hinge loss is translation invariant and $V(\boldsymbol{w})$ is the set of $\{\boldsymbol{w}_{\bar{y}}\}$ for all $\bar{y}$. The result in Sec. 3.2 easily follows.

We next consider another popular convex loss, the log-loss, given by: $\ell_g(\boldsymbol{x},y,\boldsymbol{w})=-\log\left(\frac{\exp(\boldsymbol{w}_y^T\boldsymbol{x})}{\sum_{y'}\exp(\boldsymbol{w}_{y'}^T\boldsymbol{x})}\right)$. The following corollary shows that the minimax strategy for log-loss with ℓ_2 constraints is to minimize the log-loss plus $\max_y\|\boldsymbol{w}_y\|_2$ regularization.[5]

Corollary 3.4. *The problem:*

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \max_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i)\in\mathcal{S}_{\ell_2}(\boldsymbol{x}_i;\sigma)} E_{p(\bar{\boldsymbol{x}}|\boldsymbol{x}_i)}\ell_g(\bar{\boldsymbol{x}},y_i;\boldsymbol{w}) \quad (19)$$

is equivalent to:
$\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \ell_g(\boldsymbol{x}_i,y_i,\boldsymbol{w})+\sigma\max_y\|\boldsymbol{w}_y\|_2$.

Proof. First, we note that the loss is translation invariant. Second, its subgradient is given by $-\boldsymbol{w}_y+\frac{\sum_{y'}\boldsymbol{w}_{y'}^T e^{\boldsymbol{w}_{y'}^T\boldsymbol{x}}}{\sum_{y'}e^{\boldsymbol{w}_{y'}^T\boldsymbol{x}}}$, so that $V(\boldsymbol{w})=\left\{\frac{\sum_{y'}\boldsymbol{w}_{y'}^T e^{\boldsymbol{w}_{y'}^T\boldsymbol{x}}}{\sum_{y'}e^{\boldsymbol{w}_{y'}^T\boldsymbol{x}}}\right\}_{\boldsymbol{x}\in\mathbb{R}^d}$. To see why the regularization is $\max_y\|\boldsymbol{w}_y\|_2$ note that:

$$\sup_{v\in V}\|v\|=\max_{\boldsymbol{x}}\|\frac{\sum_{y'}\boldsymbol{w}_{y'}^T e^{\boldsymbol{w}_{y'}^T\boldsymbol{x}}}{\sum_{y'}e^{\boldsymbol{w}_{y'}^T\boldsymbol{x}}}\|=\max_y\|\boldsymbol{w}_y\|_2$$

where the $\boldsymbol{x}$ that maximizes the above is equal to $\alpha\boldsymbol{w}_{\bar{y}}$ where $\bar{y}=\arg\max_y\|\boldsymbol{w}_y\|_2$ and $\alpha\to\infty$. □

Similar results may be obtained for other losses such as smoothed variants of the hinge loss, regression SVMs and others.

[5]As with the hinge loss case, in the binary classification setting this is standard ℓ_2 regularization, and otherwise it is $\ell_{2,\infty}$

Taken together, all of the above results suggest an intuitive interpretation of the regularization factor σ. Namely, it is a bound on the expected perturbation a point is likely to undergo. The regularized objective in this case is precisely the minimal worst case loss that can arise as a result of such a perturbation. This interpretation turns out to be useful in finding new ways, other then cross validation, to choose σ. One suggestion that we explore in the experiments is to relate σ to the expected distance of a point from its nearest neighbor in the training sample.

3.4 Solving RSVM$_2^2$

So far we have considered only perturbations with bounded norms. A natural question is what happens when the bound is not on a norm. A simple instance of this is when the bound is on the ℓ_2^2 of the perturbation, i.e., the perturbation set is $\mathcal{S}_{\ell_2^2}$ (see Sec. 2). As we show below, it turns out that the minimax strategy in this case is considerably more complicated, and in fact requires solving a semi-definite problem (SDP).

The following theorem states the equivalence (see the supplementary file for proof).

Theorem 3.5. *RSVM$_2^2(\sigma)$ is equivalent to the problem*

$$\min_{w_y;\alpha_i;\boldsymbol{\beta}_i;\gamma_i} \frac{1}{n}\sum_i \alpha_i\sigma + \alpha_i\|\boldsymbol{x}_i\|^2 + {\boldsymbol{x}_i}^T\boldsymbol{\beta}_i + \gamma_i$$
$$s.t.\ \forall \bar{y} \begin{bmatrix} \alpha Id & \frac{1}{2}\left(\boldsymbol{\beta}_i - \Delta^i\boldsymbol{w}_{\bar{y}}\right) \\ \frac{1}{2}\left(\boldsymbol{\beta}_i - \Delta^i\boldsymbol{w}_{\bar{y}}\right)^T & \gamma_i - e_{y_i,\bar{y}} \end{bmatrix} \succeq 0$$

Where $\Delta^i\boldsymbol{w}_{\bar{y}} = \boldsymbol{w}_{\bar{y}} - \boldsymbol{w}_{y_i}$.

The above problem has a particularly simple form in the case of binary classification as stated in the next theorem (proof in supplementary file).

Theorem 3.6. *If $y \in \{1,-1\}$, then RSVM$_2^2(\sigma)$ is equivalent to the problem*

$$\min_{\boldsymbol{w}} \frac{1}{n}\sum_i \frac{\sqrt{\sigma\|\boldsymbol{w}\|^2 + \left(1 - y\boldsymbol{w}^T\boldsymbol{x}\right)^2} + \left(1 - y\boldsymbol{w}^T\boldsymbol{x}\right)}{2}.$$

The above loss is a smoothed version of the hinge loss, with the level of smoothing determined by σ. For $\sigma = 0$ we get the hinge-loss, while for large values of σ we get the norm of $\boldsymbol{w}$ with a linear correction. Interestingly, minimizing this loss yields a consistent classifier, as was shown in a different context [19].

4 Related Work

As noted earlier, multiple works have considered minimax learning where training points are perturbed by an adversary. However, all of these [1, 12, 13, 24, 10, 8, 28] have used deterministic perturbations, where the adversary is constrained to replace a training point with a different point in the perturbation set.

Of these deterministic based approaches, the work closest to ours is [28] which shows equivalence between binary SVM and a certain minimax problem. However the model in [28] involves a rather unnatural adversary that considers the entire empirical sample jointly and splits a given perturbation budget between the training points. In our case, the adversary handles each training point independently, and this seems like a much more geometrically intuitive setting, which also results in a natural choice for σ. Furthermore, [28] need to assume non-separability of the sample, which is not needed in our case, and is often not a reasonable assumption. Finally, our analysis naturally extends to the multiclass case, and yields new and effective regularization schemes.

One work that did address stochastic noise is [3]. They studied an online setting where the *observed* feature vector is a noisy version of the true feature vector, and the noise has bounded variance. In their setting the emphasis is on recovering the true feature vector by accessing multiple instantiations of the noise. The setup is thus inherently different from ours. An interesting technical point in [3] is that they handle noise in the original feature space, even when using kernels. It would be interesting to see whether some of their tools can be used in our case to allow kernel classification with adversaries in the original space.

A complementary approach to robustness is to train a classifier that is robust to predefined or known perturbations of the training data. For example, tangent-distances [23] have been used to incorporate invariances in the training data. A similar approach was used to training SVMs [7], by expanding the training set with *virtual examples*, i.e., instances which are perturbations of existing training points, where the perturbations are obtained from *prior knowledge* of the specific problem.

Another approach to robustness is via stability of an algorithm to replacement of small subsets of the training set. It was shown theoretically [2, 18, 11, 20] that such stability analysis translates nicely to generalization bounds, and in particular support vector machines (SVMs) are stable in that sense. It would be interesting to analyze the algorithmic stability in our setting.

5 Experiments

We derived two new multiclass algorithms: RSVM$_2$ and RSVM$_2^2$. Here we compare those to standard mul-

	RSVM_2	$\text{RSVM}_2(H)$	RSVM_2^2	SVM	SVM(H)
usps	9.40	9.32	8.88	9.06	10.42
glass	47.23	52.89	50.52	48.28	56.18
ecoli	15.43	15.08	15.34	15.25	29.65
pen	7.27	11.42	7.20	7.27	15.83
poker	49.43	50.36	49.51	49.81	50.28
satimage	15.49	16.5	15.25	15.33	16.21
segment	5.65	9.18	6.15	6.25	14.31
shuttle	2.70	3.69	3.60	3.52	7.61

Figure 1: Error rates (in %) for three multiclass algorithms. The columns $\text{RSVM}_2, \text{RSVM}_2^2$ and SVM correspond to using these algorithms with cross validation for the regularization coefficient. The columns $\text{RSVM}_2(H), \text{SVM}(H)$ correspond to using the heuristic value (see text). Note that $\text{RSVM}_2(H)$ outperforms SVM(H) and is within 5% of the cross validation result RSVM_2. Result are averaged over ten random shuffles.

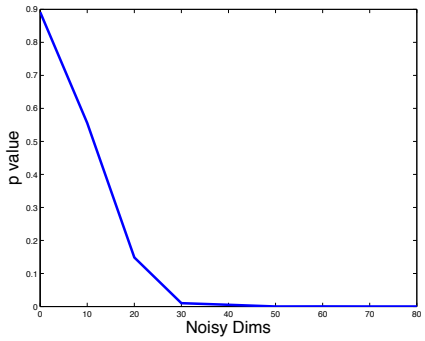

Figure 2: Comparison of $\text{RSVM}_2(\sigma)$ and SVM on toy data with noisy features added to a separable problem with 10 classes. The $p-value$ corresponding to $\text{RSVM}_2(\sigma)$ outperforming SVM is shown as a function of the number of noisy features (low values indicate that $\text{RSVM}_2(\sigma)$ is better).

ticlass SVM. We begin by considering a toy example where we have 10 classes and the original features are drawn from Gaussians in $\mathbb{R}^2$. We then add varying numbers of noisy dimensions and test the $\text{RSVM}_2(\sigma)$ and multiclass SVM (as in Eq. 2) methods. Regularization parameters are chosen by cross-validation. Fig. 2 shows the p-value for the t-test hypothesis that $\text{RSVM}_2(\sigma)$ is better than SVM (over 20 repetitions) as a function of the number of noisy dimensions. It can be seen that as more dimensions are added, $\text{RSVM}_2(\sigma)$ outperforms SVM, suggesting the former is more resilient to noise.

We next compare the performance of three methods: $\text{RSVM}_2(\sigma)$, $\text{RSVM}_2^2(\sigma)$ and multiclass SVM. Because $\text{RSVM}_2^2(\sigma)$ is relatively slower, we use training samples of up to 1000 samples.[6] We use several UCI datasets as indicated in Fig. 1. We also explore approaches to automatically choosing regularization parameters. As discussed earlier, our approach suggests a natural interpretation of the parameter in $\text{RSVM}_2(\sigma)$. Following this intuition, the chosen parameter was $\frac{1}{n\sqrt{d}}\sum_{i=1}^{n}\|x_i - N(x_i)\|_2$, where $N(x_i)$ is $\boldsymbol{x}_i$'s nearest neighbor, and d is the number of features. This reflects our interpretation of SVM whereby σ represents the expected value of the ℓ_2 perturbation a point can undergo when projected to a one dimensional space. The factor $\sqrt{d}$ reflects the fact that we are interested in the mean deviation in one particular direction. It can be derived assuming features are independent Gaussians. For comparison, in the case of $SVM(C)$ the heuristic parameter was chosen to be the default parameter choice in SVM^{light}[16], namely $\left(\frac{1}{n}\sum_{i=1}^{n} x_i^T x_i\right)^{-1}$. For the objective in eq. 2, this is given by $\frac{1}{n^2}\sum x_i^T x_i$.

We compared the above scheme for choosing σ to the common scheme of choosing it by cross validation. As shown in Fig. 1 the three algorithms yield comparable results. The heuristic parameter of $\text{RSVM}_2(\sigma)$ proves to be much better then the heuristic parameter of SVM and is within a 5% absolute difference from the cross validation result RSVM_2. This suggests that our heuristic parameter choice yields the right order of magnitude for the parameter. It can also be further improved by using a small number of cross validation search steps.

Next, we were interested in comparing the methods on large scale problems. This was done on considerably larger datasets, with over $10,000$ features and examples in each (see description in [4] and in Fig. 3). In order to run $\text{RSVM}_2(\sigma)$ on such datasets, we could not use the cvx package, but instead implemented a composite mirror descent (COMID) algorithm for the $\text{RSVM}_2(\sigma)$ objective using the approach in [9].[7] We

[6] The minimization problems $\text{RSVM}_2(\sigma)$ and $SVM(C)$ were solved using cvx [14]. $\text{RSVM}_2^2(\sigma)$ was solved using SDPT3 [25], for the inner minimization problems (i.e. the adversarial noise) and gradient descent over $\boldsymbol{w}$ using L-BFGS [17].

[7] To apply COMID to this case, one needs to solve optimization problems of the form $\min_{\boldsymbol{w}} \|\boldsymbol{w} - \boldsymbol{z}\|_2^2 + C\max_y \|\boldsymbol{w}\|_2^2$, where $\boldsymbol{z}$ is a vector and C is a constant. It turns out that this problem can be solved in closed form

	20News	Am7	Am3	EnronA	EnronB	NYTD	NYTO	NYTS	Reuters
nTraining	18,828	13,580	7,000	3,000	3,000	10,000	10,000	10,000	4000
nFeatures	252,115	686,724	494,481	13,559	18,065	108,671	108,671	114,316	23,669
nLabels	20	7	3	10	10	26	34	20	4
SVM	14.76	**23.28**	6.36	21.6	31.1	18.5	18.44	45.66	**6.53**
RSVM	**12.74**	24.37	6.36	**19.33**	**29.43**	18.31	**17.8**	**43.21**	7.38

Figure 3: Comparison of error rates (in %) between the standard multi class SVM and $\text{RSVM}_2(\sigma)$. The datasets and procedures are as in [4] (*Am7* and *Am3* stand for Amazon7 and Amazon3). Statistically significant differences are shown in bold (following the same statistical procedure as in [4]). Parameters were chosen by cross-validation. The first, second and third row show the number of training instances, the number of features and the number of labels respectively. Errors were calculated using ten fold cross validation. Bold indicates that algorithm is better with significance $p < 0.05$.

compared $\text{RSVM}_2(\sigma)$ to standard multi class SVM with Frobenius regularization [5]. The SVM was also optimized with COMID. Note that this algorithm is scalable, as it works online and the update cost is linear in the number of features and labels. Results are shown in Fig. 3. It can be seen that the two algorithms have comparable performance, with $RSVM$ being better on more datasets.

6 Discussion

We presented an analysis of minimax learning strategies where adversaries are stochastic, and have obtained several key results. First, we show that in the binary classification case, learning with bounded expected ℓ_2 norm on the perturbation is equivalent to standard SVM, where the regularization coefficient is exactly the bound on the norm.[8] This gives a natural geometric interpretation for SVM and its regularization parameter. Furthermore, it provides a natural way of choosing the regularization parameter which, as we show empirically, provides a good approximation to the error obtained via the cross-validating parameter.

Second, we show that in the multiclass case, ℓ_2 bounded perturbations are equivalent to regularization with $\ell_{2\infty}$ norm (i.e., $\max_y \|\boldsymbol{w}_y\|_2$) where the regularization parameter is again the bound on the norm. Thus, we have the same advantages as in the binary case. We note that, in retrospect, the $\ell_{2,\infty}$ makes sense as complexity regularization, perhaps even more so than the standard Frobenius norm used in this context [5]. Specifically, the $\ell_{2,\infty}$ bounds the complexity of each classifier as opposed to their sum. We believe that this should be reflected in generalization bounds (as obtained from e.g., stability, Rademacher or covering numbers) and are currently pursuing this direction.

Third, we show that the analysis can be performed for any normed perturbation on the adversary, as long as the dual norm is known and can be computed efficiently. This has interesting implications on ℓ_1 and ℓ_∞ regularization. Specifically, it implies that ℓ_1 regularized classification (e.g., [21]) can be interpreted as minimax learning with bounded ℓ_∞ perturbations. We also generalize our results to losses other than the hinge loss, and show that similar results are obtained for example for log-loss.

Finally, we show that when the perturbation is expressed in ℓ_2^2 the optimal classifier can be obtained by solving a semidefinite program. It is interesting that this small variation on the perturbation constraints results in a considerably more computationally demanding optimization problem.

Our empirical results show that there are cases where our new multiclass approach outperforms standard SVM, and that the methods are comparable on a variety of other datasets.

Many interesting questions arise from our analysis. The first is the theoretical generalization capabilities of our classifiers. As noted above, this can be approached via different tools. Other open issues are using other non-normed perturbations, and specifically structured domain dependent ones, such as translations and rotations. Another natural extension is to the structured prediction case, where interesting perturbations may also be applied to the labels themselves.

Taken together our results illustrate the utility of using stochastic adversaries for both understanding existing methods and deriving new ones.

efficiently. Complexity is linear in the number of labels and input dimension.

[8] The SVM regularizer in this case is $\|\boldsymbol{w}\|_2$ rather than $\|\boldsymbol{w}\|_2^2$ but as noted in the derivation, these are equivalent in terms of their expressive power.

Acknowledgments: This research was supported in part by a grant from the GIF, the German-Israeli Foundation for Scientific Research and Development, and in part by an Israeli Science Foundation grant ISF-1567/10. KC is a Horev Fellow, supported by the Taub Foundations.

References

[1] C. Bhattacharyya, P. Shivaswamy, and A. Smola. A second order cone programming formulation for classifying missing data. In *NIPS 17*, 2005.

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

[3] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Online learning of noisy data with kernels. In A. T. Kalai and M. Mohri, editors, *The 23rd Conference on Learning Theory*, pages 218–230. Omnipress, 2010.

[4] K. Crammer, M. Dredze, and A. Kulesza. Multi-class confidence weighted algorithms. In *EMNLP*, pages 496–504, 2009.

[5] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Machine Learning Research*, 2:265–292, March 2002.

[6] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *ACM SIGKDD*, pages 99–108, 2004.

[7] D. Decoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46(1):161–190, 2002.

[8] O. Dekel and O. Shamir. Learning to classify with missing and corrupted features. In *Proceedings of the 25th international conference on Machine learning*, pages 216–223. ACM, 2008.

[9] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In A. T. Kalai and M. Mohri, editors, *The 23rd Conference on Learning Theory*, pages 14–26, 2010.

[10] L. El Ghaoui, G. R. G. Lanckriet, and G. Natsoulis. Robust classification with interval data. Technical report, EECS Department, UC Berkeley, 2003.

[11] T. Evgeniou, M. Pontil, and A. Elisseeff. Leave one out error, stability, and generalization of voting combinations of classifiers. *Mach. Learn.*, 55:71–97, April 2004.

[12] A. Globerson and S. Roweis. Nightmare at test time: Robust learning by feature deletion. In *ICML*, pages 353–360. ACM Press, New York, NY, 2006.

[13] T. Graepel and R. Herbrich. Invariant pattern recognition by semi-definite programming machines. In *NIPS 16*. MIT Press, Cambridge, MA, 2004.

[14] M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming. *avialable at http://www. stanford. edu/boyd/cvx*, 1.

[15] K. Isii. On sharpness of Tchebycheff-type inequalities. *Annals of the Institute of Statistical Mathematics*, 14(1):185–197, 1962.

[16] T. Joachims. Making large-Scale SVM Learning Practical-Advances in Kernel Methods-Support Vector Learning, 1999.

[17] C. Kelley. *Iterative methods for linear and nonlinear equations*. Society for Industrial Mathematics, 1995.

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

[19] H. Masnadi-Shirazi and N. Vasconcelos. Variable margin losses for classifier design. In *NIPS 23*, pages 1576–1584. 2010.

[20] M. Mohri and A. Rostamizadeh. Stability bounds for non-iid processes. In *Advances in Neural Information Processing Systems*, 2007.

[21] A. Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 78–, New York, NY, USA, 2004. ACM.

[22] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[23] P. Simard, Y. LeCun, and J. S. Denker. Efficient pattern recognition using a new transformation distance. In *NIPS 5*, pages 50–58, 1993.

[24] C. Teo, A. Globerson, S. Roweis, and A. Smola. Convex learning with invariances. In *NIPS 20*, pages 1489–1496. MIT Press, Cambridge, MA, 2008.

[25] K. Toh, M. Todd, and R. Tutuncu. SDPT 3- a MATLAB software package for semidefinite programming, version 1. 3. *Optimization Methods and Software*, 11(1):545–581, 1999.

[26] L. Vandenberghe, S. Boyd, and K. Comanor. Generalized Chebyshev bounds via semidefinite programming. *SIAM review*, 49(1):52, 2007.

[27] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *Proceedings of the Seventh European Symposium on Artificial Neural Networks*, Apr. 1999.

[28] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *J. Mach. Learn. Res.*, 10:1485–1510, December 2009.