Risk Sensitive Robust Support Vector Machines

Huan Xu, Constantine Caramanis, Shie Mannor and Sungho Yun

***Abstract*— We propose a new family of classification algorithms in the spirit of support vector machines, that builds in non-conservative protection to noise and controls overfitting. Our formulation is based on a softer version of robust optimization called comprehensive robustness. We show that this formulation is equivalent to regularization by any arbitrary convex regularizer. We explain how the connection of comprehensive robustness to convex risk-measures can be used to design risk-constrained classifiers with robustness to the input distribution. Our formulations lead to easily solved convex problems. Empirical results show the promise of comprehensive robust classifiers in handling risk sensitive classification.**

I. Introduction

Support Vector Machines (SVMs) are among the most successful algorithms for classification (cf [1], [2], [3]). The standard SVM setup assumes all training samples and testing samples are independently generated according to an unknown underlying distribution, and finds a hyperplane (in the Reproducing Kernel Hilbert Space) to minimize some regularized empirical loss. In this paper we follow a different approach, proposed originally by [4], [5], [6]. The training data are assumed to be generated by the true underlying distribution, but some non-iid (potentially adversarial) disturbance is then added to the samples we observe. Previous work on robust SVMs are all based on a (often too conservative) worst-case analysis, i.e., training error under the most *adversarial* disturbance realization is considered. This worst-case approach provides a solution with but one guarantee: feasibility and worst-case performance control for *any* realization of the disturbance within the bounded uncertainty set. If the disturbance realization turns out favorable (e.g., close to mean behavior), no improved performance is guaranteed, while if the realization occurs outside the assumed uncertainty set, all bets are off: the error is not controlled. This makes it difficult to address noise with heavy tails: if one takes a small uncertainty set, there is no guarantee for high probability events; if one seeks protection over large uncertainty sets, the robust setting may yield overly pessimistic solutions.

We harness new developments in robust optimization (cf [7], [8]), in particular the softer notion of "comprehensive robustness" [9], and derive a new robust SVM formulation that addresses this problem explicitly. The key idea to comprehensive robustness is to discount lower-probability noise realizations by reducing the loss incurred. This allows us to construct classifiers with improved empirical performance together with probability bounds for *all* magnitudes of constraint violations. In particular, our contributions include: (1) We use comprehensive robustness to construct "soft robust" classifiers with performance guarantees *that depend on the level of disturbance* affecting the training data – that is, the performance guarantee is noise-level-dependent. We show that this richer class of robustness is equivalent to a much broader class of regularizers, including, e.g., standard norm-based SVM and Kullback-Leibler divergence based SVM regularizers. (2) We next show the connection to risk theory ([10], [11]), at the same time extending past work on chance constraints, and also opening the door for constructing classifiers with different risk-based guarantees. Although the connection seems natural, to the best of our knowledge this is the first attempt to view classification from a risk-hedging perspective. (3) We illustrate the performance of our new classifiers through simulation. We show that the comprehensive robust classifier, which can be viewed as a generalization of the standard SVM and the robust SVM, provides superior empirical results.

II. Comprehensive Robust Classification

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

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

where $r(\mathbf{w}, b)$ is a regularization term. The standard robust SVM (cf [5], [4]) considers the case where samples are corrupted by some noise $\vec{\boldsymbol{\delta}} = (\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_m)$ such that $\boldsymbol{\delta}_i \in \mathcal{N}_i$:

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

H. Xu and S. Mannor are with the Department of Electrical and Computer Engineering, at McGill University, Montreal, CA xuhuan@cim.mcgill.ca, shie@ece.mcgill.ca

C. Caramanis and S. Yun are with the Department of Electrical and Computer Engineering, at The University of Texas at Austin, Austin, TX caramanis@mail.utexas.edu, shyun@ece.utexas.edu

Preprint submitted to 48th IEEE Conference on Decision and Control.
Received March 6, 2009.

Let $\mathcal{N} = \prod_{i=1}^{m} \mathcal{N}_i$, and denote the hinge loss of a sample under a certain noise realization as $\xi_i(\boldsymbol{\delta}_i) \triangleq \max\big[1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b), 0\big]$. The robust classifier can be rewritten as:

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

There are two potential problems with this robust classifier. First, it treats all disturbances belonging to $\mathcal{N}$ in exactly the same manner, which can lead to an unfavorable bias to rare disturbances. In fact, it can be shown that replacing $\mathcal{N}$ with its boundary we obtain the same classifier. Second, it provides no protection against disturbances outside $\mathcal{N}$, which makes it inappropriate to handle disturbances with unbounded support, particularly in the heavy-tailed case.

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

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

We investigate additive discounts of the form $h_i(\alpha,\boldsymbol{\beta}) \triangleq \max(0, \alpha - f_i(\boldsymbol{\beta}))$ in this paper. Additive structure provides a rich class of discount functions, while remaining tractable. Moreover, additive structure provides the link to risk theory and convex risk measures which we pursue in Section IV. Substituting $h_i(\alpha,\boldsymbol{\beta}) \triangleq \max(0, \alpha - f_i(\boldsymbol{\beta}))$ and $\mathcal{N} = \prod_i \mathcal{N}_i$ into (II.1) and extending $f_i(\cdot)$ to take the value $+\infty$ for $\boldsymbol{\delta}_i \notin \mathcal{N}_i$, we obtain a formulation of the comprehensive robust classifier ($\mathbf{CRC}$):

$$\begin{aligned} \min: \quad & r(\mathbf{w},b) + \sum_{i=1}^{m} \xi_i, && \text{(II.2)}\\ \text{s.t.}: \quad & y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \ge 1 - \xi_i - f_i(\boldsymbol{\delta}_i), && \text{(II.3)}\\ & \quad \forall \boldsymbol{\delta}_i \in \mathbb{R}^n,\ i = 1,\cdots,m\\ & \xi_i \ge 0; \qquad i = 1,\cdots,m. \end{aligned}$$

Function $f_i(\cdot)$ controls the disturbance discount, and thus must satisfy $\inf_{\boldsymbol{\beta}\in\mathbb{R}^n} f_i(\boldsymbol{\beta}) = f_i(\mathbf{0}) = 0$. If we set $f_i(\cdot)$ to be the indicator function of a set, we recover the standard robust classifier. Thus the comprehensive robust classifier is a natural generalization of the robust classifier with more flexibility on setting $f_i(\cdot)$.

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

We now show that any convex regularization term in the constraint is equivalent to a comprehensive robust formulation, and vice versa. Given a function $f(\cdot)$, let f^* denote its Legendre-Fenchel transform or conjugate function, given by $f^*(s) = \sup_x\{\langle s,x\rangle - f(x)\}$. We use this below to establish the equivalence between convex regularization and comprehensive robustness. The proof of Theorem 1 is straightforward, and hence omitted.

Theorem 1 *The Comprehensive Robust Classifier (II.2) is equivalent to the convex program:*

$$\begin{aligned} \min:& r(\mathbf{w},b) + \sum_{i=1}^{m} \xi_i,\\ \text{s.t.}:& y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - f_i^*(y_i\mathbf{w}) \ge 1 - \xi_i,\ \ i = 1,\cdots,m,\\ & \xi_i \ge 0,\ i = 1,\cdots,m. \end{aligned} \quad \text{(II.4)}$$

Theorem 1 has two implications. First, it gives an equivalent and finite representation for the infinite program of the Comprehensive robust classifier. Second, the robustness for a given regularizer $f^*(\cdot)$ can be obtained by investigating the corresponding discount function $f(\cdot)$.

Since $\inf_{\mathbf{w}\in\mathbb{R}^n} f_i^*(y_i\mathbf{w}) = f_i^*(\mathbf{0}) = 0$, $f_i^*(\cdot)$ "penalizes" $y_i\mathbf{w}$ and is thus a regularization term. A classifier that has a convex regularization term $g(\cdot)$ in each constraint is equivalent to a comprehensive robust classifier with disturbance discount $f(\cdot) = g^*(\cdot)$. Therefore, the comprehensive robust classifier is equivalent to the constraint-wise regularized classifier with general convex regularization. This equivalence gives an alternative explanation for the generalization ability of regularization: the set of testing data can be regarded as a "disturbed" copy of the set of training samples where the penalty on large (or low-probability) disturbance is discounted. Empirical results show that a classifier that handles noise well has a good performance for testing samples.

As an example of this equivalence, set $f_i(\boldsymbol{\delta}_i) = \alpha\|\boldsymbol{\delta}_i\|$ for $\alpha > 0$ and $r(\mathbf{w},b) \equiv 0$. Hence, $f_i^*(y_i\mathbf{w})$ is the indicator function of the dual-norm ball with radius α. Thus (II.4) is equivalent to

$$\begin{aligned} \min:\ & \textstyle\sum_{i=1}^{m} \xi_i,\\ \text{s.t.}:\ & y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) \ge 1 - \xi_i,\ \ i = 1,\cdots,m,\\ & \|\mathbf{w}\|^* \le \alpha,\\ & \xi_i \ge 0,\ i = 1,\cdots,m. \end{aligned} \quad \text{(II.5)}$$

Problem (II.5) is the standard regularized classifier. Hence, comprehensive robust classification is a general framework which includes both robust SVMs and regularized SVMs as special cases.

Preprint submitted to 48th IEEE Conference on Decision and Control.
Received March 6, 2009.

In the full version [12], we show that as long as computing the conjugate of the discount function can be done efficiently, then the resulting comprehensive robust classification problem is computationally tractable.

III. Norm Discount

In this section, we discuss a class of discount functions based on certain ellipsoidal norms of the noise, i.e., $f_i(\boldsymbol{\delta}_i) = t_i(\|\boldsymbol{\delta}\|_V)$, for a nondecreasing $t_i : \mathbb{R}^+ \to \mathbb{R}^+$. Thus $f_i^*(\mathbf{y}) = t_i^*(\|\mathbf{y}\|_{V^{-1}})$, where $t_i^*(y) = \sup_{x\geq 0}\big[xy - t(x)\big]$. This formulation has two natural probabilistic interpretations: (1) it provides tight bounds on the probability of *all* magnitudes of constraint violations when only the first two moments of the disturbance are known (Theorem 2); (2) it computes the probabilities of *all* magnitudes of constraint violations when the disturbance is Gaussian (Theorem 3).

Theorem 2 *Suppose the random variable $\boldsymbol{\delta}_i^r$ has mean $\mathbf{0}$ and variance Σ.*[1] *Then the constraint*

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - t_i(\|\boldsymbol{\delta}_i\|_{\Sigma^{-1}}),\ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n, \tag{III.6}$$

is equivalent to

$$\inf_{\boldsymbol{\delta}_i^r\sim(0,\Sigma)} Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) \geq 1 - \frac{1}{\big(t_i^{-1}(s)\big)^2 + 1},\ \forall s \geq 0. \tag{III.7}$$

The infimum is over all 0-mean random variables with variance Σ, and $t_i^{-1}(s) \triangleq \sup\{r | t(r) \leq x\}$.

Proof: In [4], the authors study the robust formulation and show that for a fixed γ_0, the following three inequalities are equivalent:

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

Observe that Equation (III.7) is equivalent to: $\forall \gamma \geq 0$,

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

Hence, it is equivalent to:

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

Since $t_i(\cdot)$ is nondecreasing, this is equivalent to (III.6). ■

Theorem 2 shows that the comprehensive robust formulation bounds the probability of *all* magnitudes of constraint violation. Notice that the robust formulation $y_i(\langle \mathbf{w},\, \mathbf{x}_i\rangle + b) - 1 + \xi_i \geq \gamma_0\|\mathbf{w}\|_\Sigma$ gives a (tight) bound

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

explicitly depending on $\|\mathbf{w}\|_\Sigma$, and is impossible to bound without knowing $\|\mathbf{w}\|_\Sigma$ *a priori*.

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

Theorem 3 *If $\boldsymbol{\delta}_i^r \sim \mathcal{N}(0, \Sigma)$, then the constraint*

$$y_i(\langle \mathbf{w},\, \mathbf{x}_i - \boldsymbol{\delta}_i\rangle + b) \geq 1 - \xi_i - t_i(\|\boldsymbol{\delta}_i\|_{\Sigma^{-1}}),\ \forall \boldsymbol{\delta}_i \in \mathbb{R}^n, \tag{III.8}$$

is equivalent to

$$Pr\big(y_i(\langle \mathbf{w},\, \mathbf{x}_i^r\rangle + b) - 1 + \xi_i \geq -s\big) \geq \Phi\big(t_i^{-1}(s)\big), \quad \forall s \geq 0. \tag{III.9}$$

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

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

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

Notice that (III.10) is equivalent to

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

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

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

Since $t_i(\cdot)$ is nondecreasing, this is equivalent to (III.8). ■

We provide examples of "simple" discount functions. For affine function $t_i(x) = ax + b$, its conjugate function is $t_i^*(y) = I_\alpha - b$. For indicator function $t_i(x) = I_\alpha + b.$, its conjugate is $t_i(x) = ax + b$. For quadratic function $t_i(x) = ax^2 + b$, its conjugate $t_i^*(y) = y^2/4a - b$ is still a quadratic function. All these discount functions lead to second order cone programs. Another common choice of piecewise-defined discount function which replaces the jump from zero to infinity of the indicator function by a smooth increase, either linear or quadratic. Such discount functions also lead to SOCPs, and thus have a computational cost comparable to the robust formulation.

[1] We use superscript r as in $\boldsymbol{\delta}_i^r$ or $\boldsymbol{x}_i^r$ to denote the true (but unknown) value for an uncertain variable.

Preprint submitted to 48th IEEE Conference on Decision and Control.
Received March 6, 2009.

IV. Comprehensive Robustness and Convex Risk Measures

We next investigate the relationship between comprehensive robustness and convex risk measures, a notion adapted from decision theory. The theory of (convex) risk[2] measures [10] was developed in response to the observation that the preference of a decision maker among random losses (aka gambles) can be quite complicated. A risk measure defines a preference relationship among random variables: X_1 is preferable over X_2 iff $\rho(X_1) \leq \rho(X_2)$. We can regard $\rho(\cdot)$ as the measurement of how risky a random variable is: X_1 is a less risky decision than X_2 when $\rho(X_1) \leq \rho(X_2)$.

Definition 1 *A* risk measure *is a function $\rho : \mathcal{X} \to \mathbb{R}$. A risk measure is called* convex *if it satisfies the following three conditions:*

1) *Convexity:* $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$;
2) *Monotonicity:* $X \leq Y \Rightarrow \rho(X) \leq \rho(Y)$;
3) *Translation Invariance:* $\rho(X+a) = \rho(X) + a, \forall a \in \mathbb{R}$.

Convexity means diversifying reduces risk. Monotonicity says that if one random loss is always less than another, it is preferable. Translation invariance says that if a fixed penalty a is going to be paid in addition to X, we are indifferent to whether we will pay it before or after X is realized. A convex risk measure $\rho(\cdot)$ is called *normalized* if it satisfies $\rho(0) = 0$ and $\forall X \in \mathcal{X}$, $\rho(X) \geq \mathbb{E}_{\mathbb{P}}(X)$, which essentially says that the risk measure $\rho(\cdot)$ represents risk aversion. Many widely used criteria comparing random variables are normalized convex risk measures, including expected value, Conditional Value at Risk (CVaR), and the exponential loss function ([9], [13]).

Equipped with a normalized convex risk measure $\rho(\cdot)$, corresponding to particular risk preferences, we formulate the risk-measure constrained classifier ($\mathbf{RMCC}$):

$$\begin{aligned} \min: \quad & r(\mathbf{w}, b) + \sum_{i=1}^{m} \xi_i, \\ \text{s.t.}: \quad & \rho_i(\xi_i) \geq \rho_i(1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i^r \rangle + b)), \;\; i = 1, \cdots, m, \\ & \xi_i \geq 0, \;\; i = 1, \cdots, m. \end{aligned} \tag{IV.10}$$

Notice that $\mathbf{x}_i^r$ is a random variable, hence $1 - y_i(\langle \mathbf{w},\, \mathbf{x}_i^r \rangle + b)$ is a random loss, and ξ_i is the constant "equivalent." In fact, the risk-constrained classifier and the comprehensive robust classifier are equivalent.

Theorem 4 *(1) A Risk-Measure Constrained Classifier with normalized convex risk measures $\rho_i(\cdot)$ is equivalent to a Comprehensive Robust Classifier with the discount function given by*

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

(2) A Comprehensive Robust Classifier with convex discount functions $f_i(\cdot)$ is equivalent to a Risk-Constrained Classifier with the risk measure given by

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

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

Before proving Theorem 4, we establish the following two lemmas. Lemma 1 is adapted from [10], and the readers can find the proof there.

Lemma 1 *Let $\mathcal{X}$ be the set of random variables for $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{P}$ be the set of probability measures absolutely continuous with respect to $\mathbb{P}$, and $\rho : \mathcal{X} \to \mathbb{R}$ be a convex risk measure satisfying $X_n \downarrow X \Rightarrow \rho(X_n) \to \rho(X)$, then there exists a convex function $\alpha : \mathcal{P} \to (-\infty, +\infty]$ such that*

$$\rho(X) = \sup_{Q \in \mathcal{P}} \big(\mathbb{E}_Q(X) - \alpha(Q)\big) \quad \forall X \in \mathcal{X}. \tag{IV.11}$$

Furthermore, $\alpha^0(Q) \triangleq \sup_{X' \in \mathcal{X}} \big(\mathbb{E}_Q(X') - \rho(X')\big)$ satisfies (IV.11), and it is minimal in the sense that $\alpha^0(Q) \leq \alpha(Q)$ for all $Q \in \mathcal{P}$, if $\alpha(\cdot)$ also satisfies (IV.11).

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

Lemma 2 *For a normalized convex risk measure $\rho(\cdot)$, its minimal representation satisfies:*

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

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

$$\rho(0) = 0 \to \inf_{Q \in \mathcal{P}} \alpha^0(Q) = 0. \tag{IV.12}$$

Next, by definition $\alpha^0(\mathbb{P}) = \sup_{X \in \mathcal{X}} \big(\mathbb{E}_{\mathbb{P}}(X) - \rho(X)\big)$, and $\mathbb{E}_{\mathbb{P}}(X) \leq \rho(X)$ by assumption. Hence taking the supremum leads to $\alpha_0(\mathbb{P}) \leq 0$. Combining this with Equation (IV.12) establishes the lemma. ■

Now we proceed to prove Theorem 4.

Proof: (1) By Lemma 2, $f_i(\boldsymbol{\delta}_i) \geq 0$ since $\alpha^0(Q) \geq 0$, $\forall Q \in \mathcal{P}$. In addition, $\mathbb{E}_{\mathbb{P}}(\boldsymbol{\delta}_i) = \mathbf{0}$ and $\alpha^0(\mathbb{P}) = 0$ together imply $f_i(\mathbf{0}) = 0$.

Now, the constraint in the optimization formulation can be rewritten as

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

[2]This is a term used in decision theory to represent a random loss, which is different from what is often used in machine learning literature, i.e., a certain loss of the classifier.

This in turn can be rewritten as

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

which proves the first part.

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

Now, the constraint in the optimization formulation which, as above, is equivalent to

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

can also be rewritten as:

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

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

For the first part of Theorem 4, the assumption that $\rho_i(\cdot)$ is normalized can be relaxed to $\rho_i(0) = 0$ and $\inf\{\alpha_i^0(Q)|\mathbb{E}_Q(\boldsymbol{\delta}_i^r) = \mathbf{0}\} = 0$.

Let $\mathcal{P}$ be the set of probability measures absolutely continuous w.r.t. $\mathbb{P}$. It is known ([10], [11]) that any convex risk measure $\rho(\cdot)$ can be represented as $\rho(X) = \sup_{Q\in\mathcal{P}}[\mathbb{E}_Q(X) - \alpha(Q)]$ for some convex function $\alpha(\cdot)$; conversely, given any such convex function α, the resulting function $\rho(\cdot)$ is indeed a convex risk measure. Given $\alpha(\cdot)$, $\rho(\cdot)$ is called the corresponding risk measure. The function $\alpha(\cdot)$ can be thought of as a penalty function on probability distributions. This gives us a way to directly investigate classifier robustness with respect to distributional deviation. As an example, suppose we want to be robust over distributions that are nowhere more than a factor of two greater than a nominal distribution, $\mathbb{P}$. This can be captured by the risk constraint using risk measure $\rho(\cdot)$, where ρ corresponds to the convex function α given by letting $\alpha(\cdot)$ satisfy $\alpha(Q) = 0$ for $dQ/d\mathbb{P} \leq 2$, and $\alpha(Q) = +\infty$ for all other Q.

We provide some examples of classifiers obtained from such robustness w.r.t. distributional deviation.

Example 1 Suppose $\boldsymbol{\delta}_i^r \sim \mathcal{N}(0, \Sigma_i)$ and let risk measure $\rho(\cdot)$ correspond to KL-divergence,

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

Then the Risk-Measure Constrained Classifier is equivalent to

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

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

Example 2 For $\boldsymbol{\delta}_i^r$ with finite support, the Risk Measure Constrained Classifier is equivalent to

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

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

Example 3 Further let $\alpha(\mathbf{q}) = \sum_{j=1}^t q_j \log(q_j/p_j)$, the KL divergence for discrete probability measures. The risk-measure constraint in (IV.10) is equivalent to

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

This is a geometric program: it is convex, and specialized algorithms exist for its solution.

Preprint submitted to 48th IEEE Conference on Decision and Control.
Received March 6, 2009.

V. Kernelized Comprehensive Robust Classifier

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

We use $k(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ to represent the kernel function, and K to denote the Gram matrix with respect to $(\mathbf{x}_1, \cdots, \mathbf{x}_m)$. Let $\phi(\cdot)$ be the mapping from the sample space $\mathbb{R}^n$ to the feature space Φ. Let $\hat{\Phi} \subseteq \Phi$ be the subspace spanned by $\{\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_m)\}$. For a vector $\mathbf{z} \in \Phi$, denote $\mathbf{z}^=$ as its projection on $\hat{\Phi}$. The following theorem states that we can focus on $\mathbf{w} \in \hat{\Phi}$ w.l.o.g.

Theorem 5 *If $f_i(\cdot)$ is such that $f_i(\boldsymbol{\delta}) \geq f_i(\boldsymbol{\delta}^=)$, for all $\boldsymbol{\delta} \in \Phi$, and $\mathbf{w} \in \Phi$ satisfies*

$$y(\langle \mathbf{w},\, \phi(\mathbf{x}_i) - \boldsymbol{\delta}_i \rangle + b) \geq 1 - \xi_i - f_i(\boldsymbol{\delta}_i), \quad \forall \boldsymbol{\delta}_i \in \Phi, \tag{V.13}$$

then its projection $\mathbf{w}^=$ also satisfies (V.13).

Let $\mathbf{c} \triangleq (c_1, \cdots, c_m)$, $g_i(\mathbf{c}) \triangleq f_i(\sum_{i=1}^m c_i \phi(\mathbf{x}_i))$, and $\tilde{r}(\boldsymbol{\alpha}, b) \triangleq r\big(\sum_{j=1}^m \alpha_j \phi(\mathbf{x}_j), b\big)$. Let $\mathbf{e}_i$ denote the i^{th} basis vector. The kernelized comprehensive robust classifier can be written as:

Kernelized Comprehensive Robust Classifier:

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

where the constraint can be further simplified as

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

Generally $g^*(\cdot)$ depends on the exact formulation of the feature mapping $\phi(\cdot)$. However, if there exists $h_i : \mathbb{R}^+ \to \mathbb{R}^+$ such that $f_i(\boldsymbol{\delta}) = h_i(\sqrt{\langle \boldsymbol{\delta}, \boldsymbol{\delta} \rangle}), \forall \boldsymbol{\delta} \in \Phi$, then $g_i^*(y_i K\alpha) = h_i^*(\|\boldsymbol{\alpha}\|_K)$, i.e., independent of the feature mapping. When h_i is an increasing function, then $f_i(\boldsymbol{\delta}) \geq f_i(\boldsymbol{\delta}^=)$ is automatically satisfied. This is the case for many common kernels, including Gaussian RBF.

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

Theorem 6 *Suppose $h_i : \mathbb{R}^+ \to \mathbb{R}^+$ satisfies*

$$\begin{aligned} h_i\big(\sqrt{k(\mathbf{x}_i, \mathbf{x}_i) + k(\mathbf{x}_i - \boldsymbol{\delta}, \mathbf{x}_i - \boldsymbol{\delta}) - 2k(\mathbf{x}_i, \mathbf{x}_i - \boldsymbol{\delta})}\big) \\ \leq f_i(\boldsymbol{\delta}), \quad \forall \boldsymbol{\delta} \in \mathbb{R}^n. \end{aligned} \tag{V.14}$$

Then

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

implies

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

In fact, when Equation (V.14) holds with equality, this sufficient condition is also necessary. As the condition in Theorem 6 only involves the kernel function $k(\cdot,\cdot)$ and is independent of the explicit feature mapping, it applies for abstract (and infinite dimensional) mappings.

VI. Numerical Simulations

We compare the performance of three classification algorithms: the standard SVM, the standard robust SVM with ellipsoidal uncertainty set, and the comprehensive robust SVM with ellipsoidal uncertainty set with linear discount function from the center of the ellipse to its boundary (see below). The comprehensive robust classifier beats both the SVM and the robust SVM, building in protection to noise without being overly conservative.

We use a linear discount function for the comprehensive robust classifier. That is, noise is bounded in the same ellipsoidal set as for the robust SVM, $\{\boldsymbol{\delta} | \|\boldsymbol{\delta}\|_{\Sigma^{-1}} \leq 1\}$, and the discount function is $f_i(\boldsymbol{\delta}) = \alpha\|\boldsymbol{\delta}\|_{\Sigma^{-1}}$ for $\|\boldsymbol{\delta}\|_{\Sigma^{-1}} \leq 1$, and $+\infty$ otherwise. The parameter α controls the disturbance discount. As α tends to zero, there is no discount inside the uncertainty set, and we recover the robust classifier. As α tends to $+\infty$, the discount increases until the constraint is only imposed at the center of the ellipse, recovering the standard SVM.

We use SeduMi 1.1R3 ([14]) to solve the resulting convex programs. We first compare the performance of the three algorithms on the Wisconsin-Breast-Cancer data set from the UCI repository ([15]). In each iteration, we randomly pick 50% of the samples as training samples and the rest as testing samples. Each sample is corrupted by i.i.d. noise, which is uniformly distributed in an ellipsoid $\{\boldsymbol{\delta} | \|\boldsymbol{\delta}\|_{\Sigma^{-1}} \leq 1\}$. Here, the matrix Σ is diagonal. For the first 40% of features, $\Sigma_{ii} = 16$, and for the remaining features, $\Sigma_{ii} = 1$, so noise is skewed toward part of the features. We repeat 30 such iterations to get the average empirical error of the three different algorithms. Figure 1 (a) shows that the comprehensive robust classifier outperforms both the robust and standard SVM classifiers. As anticipated, when α is small, comprehensive robust classification has a testing error rate comparable to robust classification. For large α, the classifier's performance is similar to that of the standard SVM. Thus comprehensive robust classification provides a more flexible approach to handle the noise. We run similar simulations on Ionosphere and Sonar data sets from the UCI repository [15]. To fit the

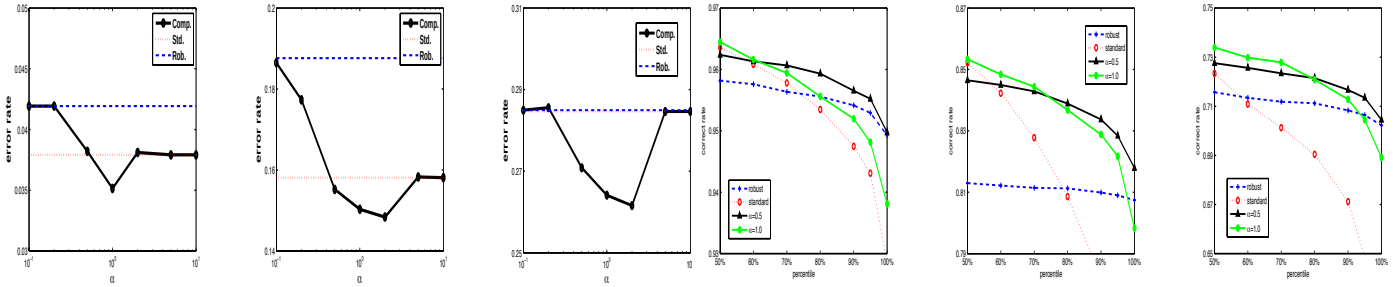

Fig. 1. Empirical Testing Error for (a) WBC Data; (b) Ionosphere Data; (c) Sonar Data.

Fig. 2. Percentile Performance for (a) WBC Data; (b) Ionosphere Data; (c) Sonar Data.

variability of the data, we scale the uncertainty set: for 40% of the features, Σ_{ii} equals 0.3 for Ionosphere and 0.01 for Sonar; for the remaining features, Σ_{ii} equals 0.0003 for Ionosphere and 0.00001 for Sonar. Figures 1 (b) and (c) show the respective simulation results. In Figure (c), the robust and standard SVM solutions coincide. As with the WBC data set, comprehensive robust classification achieves its optimal performance for mid-range α, and is superior to both the standard SVM and the robust SVM.

The noise resistance ability of the resulting classifiers is also of interest, especially when the decision maker is risk-sensitive. This is measured using percentile performance: for each testing sample, we generate 100 independent noise realizations (using the same noise model as above) and measure the probability (i.e., confidence) that this testing sample is correctly classified. The percentage of testing samples that achieves each confidence threshold is reported in Figure 2. Note that now the vertical axis represents the success rate, not the error rate. The standard SVM has a good performance for the 50% threshold, but it degrades significantly as the threshold increases, indicating a lack of noise-protection. The robust classifier tends to be overly conservative. The comprehensive robust classifier with α appropriately tuned performs well at all thresholds, especially in the 60% to 80% range, indicating good noise resistance without being overly conservative.

VII. Concluding Remarks

Our contribution is the introduction of a more geometric notion of hedging and controlling complexity (robust and comprehensive robust classifiers integrally depend on the uncertainty set and structure of the discount function) and the link to probabilistic notions of hedging, including chance constraints and convex risk constraints. We believe that the design flexibility of such a framework is the key for better performance and risk management.

References

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

[2] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 2000.

[3] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

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

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

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

[7] 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.

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

[9] A. Ben-Tal, S. Boyd, and A. Nemirovski. Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems. *Mathematical Programming, Series B*, 107:63–89, 2006.

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

[11] A. Ben-Tal, D. Bertsimas, and D. Brown. A soft robust model for optimization under ambiguity. Submitted, September 2006.

[12] H. Xu, C. Caramanis, S. Mannor, and S. Yun. Comprehensive robust support vector machines and convex risk measures - appendix. Supplementary Material, 2008.

[13] D. Bertsimas and D. B. Brown. Constructing uncertainty sets for robust linear optimization. To Appear in Operations Research, 2007.

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

[15] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

Preprint submitted to 48th IEEE Conference on Decision and Control.
Received March 6, 2009.