Mathematical Programming Series B manuscript No.
(will be inserted by the editor)

Chance Constrained Uncertain Classification via Robust Optimization

Aharon Ben-Tal · **Sahely Bhadra** ·
Chiranjib Bhattacharyya · **J. Saketha Nath**

Received: date / Accepted: date

Abstract This paper studies the problem of constructing robust classifiers when the training is plagued with uncertainty. The problem is posed as a Chance-Constrained Program (CCP) which ensures that the uncertain datapoints are classified correctly with high probability. Unfortunately such a CCP turns out to be intractable. The key novelty is in employing Bernstein bounding schemes to relax the CCP as a convex second order cone program whose solution is guaranteed to satisfy the probabilistic constraint. Prior to this work, only the Chebyshev based relaxations were exploited in learning algorithms. Bernstein bounds employ richer partial information and hence can be far less conservative than Chebyshev bounds. Due to this efficient modeling of uncertainty, the resulting classifiers achieve higher classification margins and hence better generalization. Methodologies for classifying uncertain test datapoints and error measures for evaluating classifiers robust to uncertain data are discussed. Experimental results on synthetic and real-world datasets show that the proposed classifiers are better equipped to handle data uncertainty and outperform state-of-the-art in many cases.

Keywords Chance-constraints · Bernstein Inequalities · Maximum-margin Classification · SOCP

Aharon Ben-Tal
Director, MINERVA Optimization Center, Faculty of Industrial Engg. and Management,
Technion, Haifa 32000. ISRAEL. E-mail: abental@ie.technion.ac.il,
and Extramural Fellow, *CentER*, Tilburg University, NETHERLANDS

Sahely Bhadra
Dept. of Computer Science and Automation, Indian Institute of Science, Bangalore - 12. INDIA.
E-mail: sahely@csa.iisc.ernet.in

Chiranjib Bhattacharyya
Dept. of Computer Science and Automation, Indian Institute of Science, Bangalore - 12. INDIA.
E-mail: chiru@csa.iisc.ernet.in

J. Saketha Nath
Dept. of Computer Science, Indian Institute of Technology Bombay, INDIA E-mail: saketh@cse.iitb.ac.in
Part of this work was done when the author was visiting
MINERVA Optimization Center, Faculty of Industrial Engg. and Management,
Technion, Haifa 32000. ISRAEL.

1 Introduction

Real-world classification data are fraught with uncertainty and noise. The sources of uncertainty are many — sampling errors, modeling errors or measurement errors. For example, in case of bio-medical datasets, the measurement devices vary in terms of resolutions. Lack of complete understanding of the underlying biology further complicates this problem. In case of gene/protein expression data, uncertainty is inevitable — the prime reason being biological heterogeneity of expression within the same tissue of a patient. Image classification and automated call-routing are also examples of applications where the data is prone to be erroneous.

Traditional classification algorithms, like the Support Vector Machines (SVMs) [19], assume that the training datapoints are known exactly and construct an optimal decision boundary. However, recent studies have shown that classifiers which explicitly handle the uncertainty in training data perform better than the classifiers which ignore such information [3, 13, 8]. In this paper, we propose a novel methodology for constructing maximum-margin classifiers which are robust to uncertainties in data. The proposed classifiers make no distributional assumptions regarding the underlying uncertainties and only employ partial information like support (bounds on uncertainty of the true datapoint) and second order moments (mean and variance) of the uncertain training datapoints.

In the past, robust classifiers which either employ support information [9, 4] or second order moments of the noise distribution [18] were derived. Since these classifiers employ limited partial information i.e. either support or moment information alone, though they achieve robustness to uncertainty, they tend to be overly-conservative. However, as richer partial information is employed, uncertainty can be better modeled — leading to classifiers which are robust but not overly-conservative. As discussed at various stages of this paper, a direct consequence of non-conservative modeling of the uncertainty is an increase in classification margin and hence an increase in generalizing ability of the classifier. The key contribution of this paper is to derive tractable maximum-margin formulations which employ both the support and second order moment information of the uncertain datapoints in order to build the decision boundary. Since the proposed classifiers employ richer partial information and better model the uncertainty, they achieve better generalization than the existing methods. Also, the proposed classifiers require the knowledge of bounds on second order moments rather than the exact moments, which are often unknown. Thus, in addition to being robust to uncertainty and not being overly conservative, the proposed classifiers are also inherently robust to moment estimation errors.

The idea is to derive a maximum-margin formulation which employs chance-constraints for the uncertain training datapoints. Each chance-constraint ensures that the corresponding uncertain training datapoint is classified correctly with high probability. The key novelty is to employ Bernstein bounding schemes [14, 2] for relaxing the resulting Chance-Constrained Program (CCP) as a Second Order Cone Program (SOCP), which can be efficiently solved using interior point solvers [15]. Until now only the Chebyshev bounding schemes were employed to relax various CCP based learning formulations [11, 18, 12]. To the best of our knowledge, this is the first time Bernstein approximation schemes are employed for approximately relaxing linear chance constraints via CCP based formulations. A number of alternate schemes for bounding probabilistic linear constraints exist, notably [5–7] where divergence measures other than variance are employed. It would be interesting to derive classifiers from such formulatins and will be be investigated in future. However in this paper we focus only on the Bernstein bounding based methodologies and discuss the related merits. In particular, we show that the Bernstein based schemes, by employing richer partial infor-

Table 1 Summary of formulations presented in the paper and the partial information employed by them.

S.No.	Support	1^{st} Moment	2^{nd} Moment	Formulation
1	✓	bounds	bounds	**MM-SBMV** (16), **MM-SBMV-I** (23)
2	✓	exact	exact	**MM-SMV** (24), **MM-SMV-I** (26)
3	✓	bounds	×	**MM-SBM** (27), **MM-SBM-I** (29)
4	✓	exact	×	**MM-SM** (30), **MM-SM-I** (32)

mation (support and second order moment information), lead to less conservative modeling of the uncertainty than the Chebyshev based schemes, which employ moment information alone. Using this SOCP relaxation as a basis, various maximum-margin formulations are derived which employ different levels of information about the uncertain datapoints. Table 1 summarizes the formulations[1] derived in the paper and the partial information employed by them.

The remainder of the paper is organized as follows: in section 1.1, the past work done on maximum-margin classification with uncertain data is briefly reviewed. Section 2 presents the main contribution of the paper, a maximum-margin SOCP formulation which employs the support and bounds on the second order moments of the uncertain datapoints in order to achieve robustness. The section also presents various specializations of this formulation to the scenarios presented in table 1.The subsequent section discusses the issue of classifying uncertain test datapoints and presents various error measures which evaluate the performance of classifiers which handle uncertain data. In section 4, experimental results which compare the performance of the proposed methods and the existing methods are presented. The paper concludes in section 5, by summarizing the work.

1.1 Review of Past Work

In this section, we review the work done on maximum-margin classification with uncertain data. We start by discussing the well known SVM formulation [19], which assumes that the training datapoints are known exactly. Here, a hyperplane, $\mathbf{w}^\top \mathbf{x} - b = 0$, that maximally separates the positive and negative training datapoints is constructed. Denoting the training datapoints by $X_i \equiv [X_{i1} \dots X_{in}]^\top \in \mathbb{R}^n,\ i = 1, \dots, m$ and the respective class labels by $y_i,\ i = 1, \dots, m$, this problem can be expressed as:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi_i} \quad & \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\textstyle\sum_{i=1}^m \xi_i \\ \text{s.t. } & y_i(\mathbf{w}^\top X_i - b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \dots, m \end{aligned} \tag{1}$$

ξ_i are slack variables introduced to allow for outliers and C is a user-given regularization parameter. Note that the objective minimizes $\|\mathbf{w}\|_2$, which turns out to be inversely proportional to the margin of separation achieved between the positive and negative datapoints. According to the structural risk minimization principle of Vapnik [19], such classifiers which maximize the margin achieve good generalization.

[1] Nomenclature of formulations: prefix "**MM**" denotes **M**aximum **M**argin classifier. Partial information of **S**upport, **M**ean, **V**ariance employed by the classifier are denoted by '**S**', '**M**', '**V**' respectively. The symbol '**B**' denotes that the corresponding classifier employs bounds on moments rather than exact moments. The suffix '**I**' indicates that the corresponding classifier is a variant, whose meaning will be clear later in the text. For e.g., the abbreviation "**MM-SBMV**" stands for a maximum-margin classifier which employs support, bounds on means and variances of uncertain datapoints.

However, if the training datapoints, X_i, are known to be uncertain and information regarding the underlying uncertainties is provided, then classifiers which utilize such information generalize better than their non robust counterparts (e.g. SVMs [18,9]). Different approaches assume different kinds of information regarding the uncertainties is known. The simplest of these is a maximum-margin classifier which employs just the means of the uncertain datapoints, $\mu_i \equiv \mathbf{E}[X_i]$. The problem solved is then:

$$\textbf{(MM-M)} \qquad \min_{\mathbf{w},b,\xi_i} \quad \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\textstyle\sum_{i=1}^m \xi_i$$
$$\text{s.t. } y_i(\mathbf{w}^\top \mu_i - b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1,\ldots,m \qquad (2)$$

Assuming uncertainty in the datapoints is bounded i.e., the support is known, tractable classification formulations which are robust to uncertainty can been derived [9,4]. Specifically, [9] assume that the extremum values of the features of the datapoints are known i.e., $l_{ij} \leq X_{ij} \leq u_{ij}$. In other words, each training datapoint X_i is assumed to lie in a hyper-rectangle: $\mathscr{R}_i \equiv \left\{\mathbf{x} = [x_1 \ldots x_n]^\top \in \mathbb{R}^n \mid l_{ij} \leq x_j \leq u_{ij},\ j = 1,\ldots,n\right\}$ and constraints enforcing correct classification of all the datapoints lying in a bounding hyper-rectangle are imposed: $y_i(\mathbf{w}^\top \mathbf{x} - b) \geq 1 - \xi_i,\ \forall\, \mathbf{x} \in \mathscr{R}_i$. This leads to the following problem:

$$\textbf{(MM-S)} \qquad \min_{\mathbf{w},b,\xi_i} \quad \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\textstyle\sum_{i=1}^m \xi_i$$
$$\text{s.t. } y_i(\mathbf{w}^\top \mathbf{c}_i - b) \geq 1 - \xi_i + \|\mathbf{S}_i\mathbf{w}\|_1,\ \xi_i \geq 0,\ i = 1,\ldots,m \qquad (3)$$

where $\mathbf{c}_i$ is the geometric center of the hyper-rectangle $\mathscr{R}_i$ and $\mathbf{S}_i$ is a diagonal matrix with entries as semi-lengths of the sides of the hyper-rectangle $\mathscr{R}_i$. Using the means, μ_i, and covariances, $\Sigma_i \equiv \mathbf{cov}[X_i]$ of the uncertain training datapoints, and employing the Chebyshev's inequality, classifiers which are robust to uncertainty have been derived [3,18]:

$$\textbf{(MM-MC)} \qquad \min_{\mathbf{w},b,\xi_i} \quad \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\textstyle\sum_{i=1}^m \xi_i$$
$$\text{s.t. } y_i(\mathbf{w}^\top \mu_i - b) \geq 1 - \xi_i + \kappa_c\|\Sigma_i^{\frac{1}{2}}\mathbf{w}\|_2,\ \xi_i \geq 0,\ i = 1,\ldots,m \qquad (4)$$

where $\kappa_c = \sqrt{\frac{1-\varepsilon}{\varepsilon}}$ and $\varepsilon \in [0,1]$ is a user-given parameter. The robust formulations derived in [4] turn out to be special cases of the (**MM-MC**) formulation.

Each of the three robust formulations presented above differ in the way uncertainty is modeled using various partial information like support, first and second order moments. The formulation (**MM-M**) uses only mean (first order moment) information, while (**MM-S**) uses support information and (**MM-MC**) uses second order moment information. The conservative nature of a formulation depends on the partial information employed by it. As more information is employed, the uncertainty can be better modeled — leading to robust as well as non-overly-conservative classifiers. Now, the conservative nature of a robust classifier has direct influence over the generalization ability of the classifier — this is justified in the following text. Note that, more conservative the uncertainty modeling is, tighter are the classification constraints in the respective formulations. For example, (**MM-S**) models the uncertain datapoint using its bounding hyper-rectangle whereas (**MM-M**) models it as the single point μ_i. Clearly, the classification constraints (3) in (**MM-S**) which imply that the entire hyper-rectangle must be classified correctly are tighter than those in (**MM-M**) which imply that the mean, μ_i, alone needs to be classified correctly. It is also easy to see that because of this conservative modeling of uncertainty in (**MM-S**), the margin, $\frac{1}{2}\|\mathbf{w}\|_2^2$, achieved

by it is lesser than that with (**MM-M**). According to the structural risk minimization principle [19], larger is the margin of a classifier, better is its generalization ability. Thus (**MM-S**), though robust to uncertainty fails to generalize well due to its conservative nature. On the other hand, (**MM-M**), though models uncertainty in a less conservative manner, it is not robust enough as it assumes mean is the only possible position for the uncertain datapoint. Thus in order to achieve good generalization classifiers need to be robust to uncertainties in data while not being overly-conservative.

The formulation (**MM-MC**) is nearest in spirit to the present work. As shown in [18], (**MM-MC**) is the result of relaxing a CCP based formulation using the Chebyshev inequality. Relaxation schemes based on the Chebyshev's inequality are known to be conservative as they employ second moment information alone. In this paper, we employ Bernstein bounding schemes in order to relax the same CCP based maximum-margin formulation. The Bernstein based relaxation employs both the support and second order moment information and hence leads to less conservative modeling of the uncertainty, which as discussed above is key in deriving classifiers with good generalization.

2 Maximum-margin Formulations for Uncertain Data

This section presents the novel maximum-margin classification formulations which are robust to uncertainties in training data. The notation used is summarized below: let $X_i \equiv [X_{i1} \dots X_{in}]^\top$ be the random variable generating the i^{th} (uncertain) training datapoint and let y_i be its label. The following information regarding the uncertain datapoints is assumed to be known:

Support Extremum values of features of the datapoints are known i.e., $l_{ij} \leq X_{ij} \leq u_{ij}$. In other words, $X_i \in \mathscr{R}_i \equiv \left\{\mathbf{x} = [x_1 \dots x_n]^\top \in \mathbb{R}^n \mid l_{ij} \leq x_j \leq u_{ij},\ j = 1, \dots, n\right\}$.

1st Moment Bounds on the means of the datapoints, $\mu_i^- \equiv [\mu_{i1}^- \dots \mu_{in}^-]^\top \leq \mu_i \equiv [\mu_{i1} \dots \mu_{in}]^\top = \mathbf{E}[X_i] \equiv [\mathbf{E}[X_{i1}] \dots \mathbf{E}[X_{in}]]^\top \leq \mu_i^+ \equiv [\mu_{i1}^+ \dots \mu_{in}^+]^\top$.

2nd Moment Bounds on second-moments of the feature values of the datapoints are known i.e. $0 \leq \mathbf{E}[X_{ij}^2] \leq \sigma_{ij}^2$.

Note that no assumptions regarding the forms of the uncertainty distributions are made. The discriminating hyperplane which is to be learnt using the given training data is denoted by $\mathbf{w}^\top \mathbf{x} - b = 0$, where $\mathbf{w} \equiv [w_1 \dots w_n]^\top$ is the normal and b is the bias of the hyperplane. Recall the SVM formulation (1), which we consider here as the baseline formulation. Now, since the datapoints X_i are uncertain, the constraints in (1) can no longer be satisfied always. Hence, alternatively, it is required that the following chance-constraints are satisfied:

$$Prob\left(y_i(\mathbf{w}^\top X_i - b) \leq 1 - \xi_i\right) \leq \varepsilon, \tag{5}$$

where $0 \leq \varepsilon \leq 1$ is a user-given parameter close to 0, denoting an upper bound on the mis-classification error made on X_i. Thus, the chance-constraints in (5) ensure that the uncertain datapoints are mis-classified with small probability. Using these chance-constraints, the following maximum-margin formulation, similar in spirit to SVMs, can be written:

$$\textbf{(CCP)} \quad \min_{\mathbf{w},b,\xi_i} \quad \tfrac{1}{2}\|\mathbf{w}\|_2^2 + C\textstyle\sum_{i=1}^m \xi_i$$
$$\text{s.t. } Prob\left(y_i(\mathbf{w}^\top X_i - b) \leq 1 - \xi_i\right) \leq \varepsilon,\ \xi_i \geq 0,\ i = 1, \dots, m \tag{6}$$

The above formulation (**CCP**) is hard to solve even when the probability distribution of the X_i's are fully known, because the constraints are typically non-convex. In the remainder of the section several safe convex approximations of (**CCP**) are derived, assuming different levels of partial information regarding the uncertainties are known.

2.1 Formulations using Support and Bounds on 2^{nd} Order Moments

In this section, we present a maximum-margin classification formulation which employs bounds on means and variances, as well as support (bounding hyper-rectangles) of the uncertain training datapoints are known. It is also assumed that the features used to describe the data are independent — in other words, the random variables $X_{ij},\ j = 1,\ldots,n$ are assumed to be independent. The key idea is to derive convex constraints involving the above partial information, which when satisfied imply that the chance-constraints (5) are satisfied. To this end, the following theorem is presented, which specializes the Bernstein approximation schemes described in [14, 2, 1]:

Theorem 1 *Assuming partial information of support ($l_{ij} \leq X_{ij} \leq u_{ij}$), bounds on first-moments ($\mu_{ij}^- \leq \mu_{ij} = \mathbf{E}[X_{ij}] \leq \mu_{ij}^+$) and bounds on second-moments ($0 \leq \mathbf{E}[X_{ij}^2] \leq \sigma_{ij}^2$) of independent random variables $X_{ij},\ j = 1,\ldots,n$ are known, the chance-constraint (5) is satisfied if the following **convex** constraint in variables, $(\mathbf{w}, b, \xi_i)$, holds:*

$$1 - \xi_i + y_i b + \sum_j \left(\max\left[-y_i \mu_{ij}^- w_j, -y_i \mu_{ij}^+ w_j \right] \right) + \kappa \|\Sigma_{(1),i} \mathbf{w}\|_2 \leq 0 \tag{7}$$

where $\kappa \equiv \sqrt{2\log(1/\varepsilon)}$, $\Sigma_{(1),i}$ is a diagonal matrix given by:

$$\Sigma_{(1),i} = diag\left(\left[s_{i1} \nu\left(\mu_{i1}^-, \mu_{i1}^+, \sigma_{i1}\right) \ldots s_{in} \nu\left(\mu_{in}^-, \mu_{in}^+, \sigma_{in}\right) \right] \right), \tag{8}$$

$s_{ij} \equiv \frac{u_{ij} - l_{ij}}{2}$ and the function $\nu\left(\mu_{ij}^-, \mu_{ij}^+, \sigma_{ij}\right)$ is as defined in (14).

Proof The chance-constraint (5) can be written as:

$$Prob\left(\mathbf{a}_i^\top X_i + a_{i0} \geq 0 \right) \leq \varepsilon \tag{9}$$

where $a_{i0} = 1 - \xi_i + y_i b$ and $\mathbf{a}_i = -y_i \mathbf{w}$.

Using Markov inequality and independence of random variables, $X_{ij}, j = 1,\ldots,n$, we have that:

$$Prob\left(\mathbf{a}_i^\top X_i + a_{i0} \geq 0 \right) \leq \exp\{\alpha a_{i0}\} \prod_j \mathbf{E}\left[\exp\{\alpha a_{ij} X_{ij}\}\right],\ \forall\ \alpha \geq 0 \tag{10}$$

Key to modeling chance constraint (9) now depends on how one upperbounds the moment generating functions, $\mathbf{E}\left[\exp\{tX_{ij}\}\right],\ t \in \mathbb{R}$. To continue the proof, we use the following lemma:

Lemma 1 *Suppose the support and bounds on first, second moments of the random variable X_{ij} are known. Then,*

$$\mathbf{E}\left[\exp\{tX_{ij}\}\right] \leq \exp\left\{ \frac{\nu\left(\mu_{ij}^-, \mu_{ij}^+, \sigma_{ij}\right)^2 s_{ij}^2}{2} t^2 + \max\left[\mu_{ij}^- t, \mu_{ij}^+ t\right] \right\}\ \forall\, t\ \in\ \mathbb{R} \tag{11}$$

Proof Consider the normalized random variable $\hat{X}_{ij} \equiv \frac{X_{ij}-c_{ij}}{s_{ij}}$, where $c_{ij} \equiv \frac{l_{ij}+u_{ij}}{2}$ and $s_{ij} \equiv \frac{u_{ij}-l_{ij}}{2}$. It is easy to see that $-1 \leq \hat{X}_{ij} \leq 1$, $\mathbf{E}[\hat{X}_{ij}] = \frac{\mathbf{E}[X_{ij}]-c_{ij}}{s_{ij}}$ and $\mathbf{E}[\hat{X}_{ij}^2] = \frac{\mathbf{E}[X_{ij}^2]-2\mathbf{E}[X_{ij}]c_{ij}+c_{ij}^2}{s_{ij}^2}$. Using these relations one can easily compute the bounds on first and second moments of $\hat{X}_{ij}$. Let these be denoted by $\hat{\mu}_{ij}^- \leq \hat{\mu}_{ij} = \mathbf{E}[\hat{X}_{ij}] \leq \hat{\mu}_{ij}^+$ and $0 \leq \mathbf{E}[\hat{X}_{ij}^2] \leq \hat{\sigma}_{ij}^2$ respectively. By Jensen's inequality, we have that $|\hat{\mu}_{ij}| \leq \hat{\sigma}_{ij}$. Hence, without loss of generality, assume $|\hat{\mu}_{ij}^{\pm}| \leq \hat{\sigma}_{ij}$. Now, $\mathbf{E}\left[\exp\{tX_{ij}\}\right] = \mathbf{E}\left[\exp\{ts_{ij}\hat{X}_{ij}\}\right]\exp\left\{tc_{ij}\right\}$. Let $\tilde{t} \equiv ts_{ij}$. We know that (refer table 2 in [14], chapter 2 in [1]):

$$\mathbf{E}\left[\exp\{\tilde{t}\hat{X}_{ij}\}\right] \leq g_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t}) \equiv \begin{cases} \frac{\left(1-\hat{\mu}_{ij}\right)^2 \exp\left\{\tilde{t}\frac{\hat{\mu}_{ij}-\hat{\sigma}_{ij}^2}{1-\hat{\mu}_{ij}}\right\}+\left(\hat{\sigma}_{ij}^2-\hat{\mu}_{ij}^2\right)\exp\{\tilde{t}\}}{1-2\hat{\mu}_{ij}+\hat{\sigma}_{ij}^2}, & \tilde{t} \geq 0 \\ \frac{\left(1+\hat{\mu}_{ij}\right)^2 \exp\left\{\tilde{t}\frac{\hat{\mu}_{ij}+\hat{\sigma}_{ij}^2}{1+\hat{\mu}_{ij}}\right\}+\left(\hat{\sigma}_{ij}^2-\hat{\mu}_{ij}^2\right)\exp\{-\tilde{t}\}}{1+2\hat{\mu}_{ij}+\hat{\sigma}_{ij}^2}, & \tilde{t} \leq 0 \end{cases} \tag{12}$$

Note that the above bound is tight given the circumstances: for $t > 0$, the bound is achieved by a 2-point distribution at the points $\frac{\hat{\mu}_{ij}-\hat{\sigma}_{ij}^2}{1-\hat{\mu}_{ij}}$ and 1 with masses $\frac{(1-\hat{\mu}_{ij})^2}{1-2\hat{\mu}_{ij}+\hat{\sigma}_{ij}^2}$ and $\frac{\hat{\sigma}_{ij}^2-\hat{\mu}_{ij}^2}{1-2\hat{\mu}_{ij}+\hat{\sigma}_{ij}^2}$ respectively. For such a distribution, the mean is indeed $\hat{\mu}_{ij}$ and the second moment is $\hat{\sigma}_{ij}^2$. Similar arguments hold for the case $t < 0$. Though the bound in (12) is the tightest possible under the given circumstances, employing it in (10) will not lead to tractable relaxations of the original chance-constraint. Hence we further upper bound the RHS of (12) by a single exponential function such that the final relaxed constraint is tractable. To this end, define the function:

$$h_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t}) \equiv \log g_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t}) \tag{13}$$

It is easy to show that $h_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(0) = 0$ and $h'_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(0) = \hat{\mu}_{ij}$. Now for $\tilde{t} \geq 0$,

$$\begin{aligned} h''_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t}) &= \frac{\left(\hat{\sigma}_{ij}^2-\hat{\mu}_{ij}^2\right)\left(1-2\hat{\mu}_{ij}+\hat{\sigma}_{ij}^2\right)^2 \exp\left\{\tilde{t}\frac{\hat{\mu}_{ij}-\hat{\sigma}_{ij}^2}{1-\hat{\mu}_{ij}}\right\}\exp\{\tilde{t}\}}{\left[(1-\hat{\mu}_{ij})^2 \exp\left\{\tilde{t}\frac{\hat{\mu}_{ij}-\hat{\sigma}_{ij}^2}{1-\hat{\mu}_{ij}}\right\}+\left(\hat{\sigma}_{ij}^2-\hat{\mu}_{ij}^2\right)\exp\{\tilde{t}\}\right]^2} \\ &\leq \frac{4\left(\hat{\sigma}_{ij}^2-\hat{\mu}_{ij}^2\right)(1-\hat{\mu}_{ij})^2 \exp\left\{\tilde{t}\frac{\hat{\mu}_{ij}-\hat{\sigma}_{ij}^2}{1-\hat{\mu}_{ij}}\right\}\exp\{\tilde{t}\}}{\left[(1-\hat{\mu}_{ij})^2 \exp\left\{\tilde{t}\frac{\hat{\mu}_{ij}-\hat{\sigma}_{ij}^2}{1-\hat{\mu}_{ij}}\right\}+\left(\hat{\sigma}_{ij}^2-\hat{\mu}_{ij}^2\right)\exp\{\tilde{t}\}\right]^2} \leq 1 \end{aligned}$$

The last inequality is true by the AM-GM inequality . Similarly one can derive an inequality for the case $\tilde{t} \leq 0$. Thus $h''_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t}) \leq 1 \ \forall \ \tilde{t}$. Using Taylor series, it follows that $h_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t}) \leq \hat{\mu}_{ij}\tilde{t} + \frac{1}{2}\tilde{t}^2 \ \forall \tilde{t}$. As a result, the function:

$$\nu\left(\mu^-,\mu^+,\sigma\right) \equiv \min\left\{k \geq 0 \ : \ h_{\hat{\mu},\hat{\sigma}}(\tilde{t}) \leq \max[\hat{\mu}^-\tilde{t},\hat{\mu}^+\tilde{t}] + \frac{k^2}{2}\tilde{t}^2 \ \forall \ (\hat{\mu} \in [\hat{\mu}^-,\hat{\mu}^+],\tilde{t})\right\} \tag{14}$$

is well defined (in fact $0 \leq \nu(\cdot,\cdot,\cdot) \leq 1$) and hence

$$g_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t}) \leq \exp\left\{\max[\hat{\mu}_{ij}^-\tilde{t},\hat{\mu}_{ij}^+\tilde{t}] + \frac{\nu\left(\mu_{ij}^-,\mu_{ij}^+,\sigma_{ij}\right)^2}{2}\tilde{t}^2\right\} \ \forall \tilde{t}$$

Noting that $g_{\hat{\mu}_{ij},\hat{\sigma}_{ij}}(\tilde{t})$ is an upper bound on $\mathbf{E}\left[\exp\{\tilde{t}\hat{X}_{ij}\}\right]$ and using the fact that $\mathbf{E}\left[\exp\{tX_{ij}\}\right] = \mathbf{E}\left[\exp\{ts_{ij}\hat{X}_{ij}\}\right]\exp\left\{tc_{ij}\right\}$ and $\mu_{ij}^{\pm} = s_{ij}\hat{\mu}_{ij}^{\pm} + c_{ij}$, we obtain (11). This completes the proof of Lemma 1. □

Using lemma 1 and (10) we obtain ($\forall\ \alpha \geq 0$):

$$\log\left[Prob\left(\mathbf{a}_i^{\top}X_i + a_{i0} \geq 0\right)\right] \leq \alpha\left(a_{i0} + \sum_j\left(\max\left[-y_i\mu_{ij}^{-}w_j, -y_i\mu_{ij}^{+}w_j\right]\right)\right) + \frac{\alpha^2}{2}\|\Sigma_{(1),i}\mathbf{w}\|_2^2$$

Since this inequality holds for all non-negative α's, if we ensure that for certain α the right-hand side of the inequality is $\leq \log(\varepsilon)$, then we would satisfy the chance-constraint (9). So, we have:

$$\underbrace{\alpha\left(a_{i0} + \sum_j\left(\max\left[-y_i\mu_{ij}^{-}w_j, -y_i\mu_{ij}^{+}w_j\right]\right)\right)}_{p} + \frac{\alpha^2}{2}\underbrace{\|\Sigma_{(1),i}\mathbf{w}\|_2^2}_{q^2} \leq \log\varepsilon \tag{15}$$

In the case $q = 0$, the above inequality is possible only if $p < 0$ ($\because \varepsilon \in [0,1]$). Now suppose $q > 0$. We wish to choose that value of α for which the LHS of (15) is minimized. This minimized value is 0 if $p \geq 0$ and $-\frac{p^2}{2q^2}$ if $p < 0$. Again since $\varepsilon \in [0,1]$, $p \geq 0$ is not allowed. Substituting $-\frac{p^2}{2q^2}$ in LHS of (15), we have $\frac{p^2}{q^2} \leq \kappa^2 \Leftrightarrow p + \kappa q \leq 0$ ($\because p < 0$). Hence either in the case $q = 0$ or $q > 0$, $p + \kappa q \leq 0$ is the sufficient condition for satisfying the chance-constraint (5). Substituting the values of p, q, a_{i0} in this inequality we obtain (7). This completes the proof of Theorem 1. □

Replacing the chance-constraints (6) in (**CCP**) with the deterministic (convex) constraints (7), we obtain a maximum-margin formulation which ensures that the probability of misclassification when trained with uncertain data, X_i, is less than ε. This formulation can be written as the following SOCP:

$$\begin{array}{lll} \textbf{(MM-SBMV)} & \min\limits_{\mathbf{w},b,\xi_i,z_{ij}} & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum\limits_{i=1}^{m}\xi_i \\ & \text{s.t.} & 1 - \xi_i + y_i b + \sum\limits_j z_{ij} + \kappa\|\Sigma_{(1),i}\mathbf{w}\|_2 \leq 0, \\ & & z_{ij} \geq -y_i\mu_{ij}^{-}w_j,\ z_{ij} \geq -y_i\mu_{ij}^{+}w_j,\ \xi_i \geq 0 \end{array} \tag{16}$$

The values of the function $\nu(\mu_{ij}^{-}, \mu_{ij}^{+}, \sigma_{ij})$ (14) can be calculated numerically. The details of the numerical procedure are presented in section 2.1.1. The SOCP (**MM-SBMV**) can be efficiently solved using cone program solvers like `SeDuMi`[2], `Mosek`[3] or `CPLEX`[4].

In the following, a geometrical interpretation of the formulation (**MM-SBMV**) is presented. To this end, consider the following lemma:

[2] Available at `http://sedumi.mcmaster.ca/`

[3] Available at `http://www.mosek.com/index.php?id=7`

[4] Available at `http://www.ilog.com/products/cplex/`

Lemma 2 *Let the set*

$$\mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right) \equiv \left\{\mathbf{x} = \mu_i + \kappa\Sigma_{(1),i}\mathbf{a} \,:\, \|\mathbf{a}\|_2 \leq 1\right\} \tag{17}$$

represent an ellipsoid centered at μ_i, whose shape and size are determined by $\kappa\Sigma_{(1),i}$. Consider the problem of correctly classifying points belonging to the union of ellipsoids $\mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right)$ over $\mu_i \in [\mu_i^-, \mu_i^+]$:

$$y_i(\mathbf{w}^\top\mathbf{x} - b) \geq 1 - \xi_i,\ \forall\, \mathbf{x} \in \cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right) \tag{18}$$

The continuum of constraints (18) are satisfied if and only if (7) holds.

Proof We have the following:

$$\begin{aligned}
(18) &\Leftrightarrow \max_{\mathbf{x} \in \cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right)} \left(-y_i\mathbf{w}^\top\mathbf{x}\right) + 1 - \xi_i + y_i b \leq 0 \\
&\Leftrightarrow \max_{\mu_i \in [\mu_i^-, \mu_i^+], \|\mathbf{a}\|_2 \leq 1} \left(-y_i\mathbf{w}^\top(\mu_i + \kappa\Sigma_{(1),i}\mathbf{a})\right) + 1 - \xi_i + y_i b \leq 0 \\
&\Leftrightarrow \max_{\mu_i \in [\mu_i^-, \mu_i^+]} \left(-y_i\mathbf{w}^\top\mu_i\right) + \max_{\|\mathbf{a}\|_2 \leq 1} \left(-\kappa y_i\mathbf{w}^\top\Sigma_{(1),i}\mathbf{a}\right) + 1 - \xi_i + y_i b \leq 0 \\
&\Leftrightarrow (7)
\end{aligned}$$

This completes the proof. □

The above lemma shows that the formulation (**MM-SBMV**) views each uncertain training datapoint as the set $\cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right)$ and does a maximum-margin classification using these uncertainty sets.

Note that the size of uncertainty set, and hence robustness and conservative nature of the classifier depend on κ (and hence on ε). More specifically, as the upper bound on misclassification error, ε, decreases, size of the uncertainty set increases. However from the support information we know that the true training datapoint can never lie outside its bounding hyper-rectangle. Thus we can obtain less conservative classifiers by employing constraints using uncertainty sets as the intersection of $\cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right)$ and the bounding hyper-rectangle $\mathscr{R}_i$. To this end we present the following lemma:

Lemma 3 *Consider the problem of correctly classifying points belonging to the set $\mathscr{R}_i \cap \left(\cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right)\right)$:*

$$y_i(\mathbf{w}^\top\mathbf{x} - b) \geq 1 - \xi_i,\ \forall\, \mathbf{x} \in \mathscr{R}_i \cap \left(\cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(1),i}\right)\right) \tag{19}$$

*The continuum of constraints (19) are satisfied if and only if the following **convex** constraint in $(\mathbf{w}, b, \xi_i, \mathbf{a}_i)$ holds (here, $\mathbf{a}_i \equiv [a_{i1} \ldots a_{in}]^\top$):*

$$1 - \xi_i + y_i b + \sum_j \left(\max\left[-l_{ij}\left(y_i w_j + a_{ij}\right), -u_{ij}\left(y_i w_j + a_{ij}\right)\right] + \max\left[\mu_{ij}^- a_{ij}, \mu_{ij}^+ a_{ij}\right]\right) + \kappa\|\Sigma_{(1),i}\mathbf{a}_i\|_2 \leq 0 \tag{20}$$

Proof The constraints (19) hold if and only if:

$$1-\xi_i+y_ib+\left(\max_{\mathbf{x}\,\in\,\mathscr{R}_i\cap\left(\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)\right)} -y_i\mathbf{w}^\top\mathbf{x}\right)\leq 0$$

Note that, the term with max in the above inequality is nothing but the support function of the set $\mathscr{R}_i\cap\left(\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)\right)$, denoted by $I_{\mathscr{R}_i\cap\left(\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)\right)}(-y_i\mathbf{w})$. Since support function of intersection of two sets is the infimal convolution of support functions of the individual sets (see section 16, [16]), we have that:

$$\begin{aligned}(19)\Leftrightarrow\ & 1-\xi_i+y_ib+\inf_{\mathbf{a}_i+\bar{\mathbf{a}}_i=-y_i\mathbf{w}}\left\{I_{\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)}(\mathbf{a}_i)+I_{\mathscr{R}_i}(\bar{\mathbf{a}}_i)\right\}\leq 0\\ \Leftrightarrow\ & \exists\,\mathbf{a}_i,\bar{\mathbf{a}}_i\ni\ 1-\xi_i+y_ib+\left\{I_{\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)}(\mathbf{a}_i)+I_{\mathscr{R}_i}(\bar{\mathbf{a}}_i)\right\}\leq 0,\ \mathbf{a}_i+\bar{\mathbf{a}}_i=-y_i\mathbf{w}\end{aligned}\tag{21}$$

Let the entries in vectors $\mathbf{a}_i,\bar{\mathbf{a}}_i$ be $a_{ij},\bar{a}_{ij},\ j=1,\ldots,n$ respectively. Then by lemma 2, we have that $I_{\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)}(\mathbf{a}_i)=\sum_j\left(\max\left[\mu_{ij}^-a_{ij},\mu_{ij}^+a_{ij}\right]\right)+\kappa\|\Sigma_{(1),i}\mathbf{a}_i\|_2$. Also, $I_{\mathscr{R}_i}(\bar{\mathbf{a}}_i)=\sum_j\max\left[l_{ij}\bar{a}_{ij},u_{ij}\bar{a}_{ij}\right]$. Hence, we have that (19) is satisfied if and only if:

$$1-\xi_i+y_ib+\sum_j\max\left[l_{ij}\bar{a}_{ij},u_{ij}\bar{a}_{ij}\right]+\sum_j\left(\max\left[\mu_{ij}^-a_{ij},\mu_{ij}^+a_{ij}\right]\right)+\kappa\|\Sigma_{(1),i}\mathbf{a}_i\|_2\leq 0\tag{22}$$

and $\mathbf{a}_i+\bar{\mathbf{a}}_i=-y_i\mathbf{w}$. Eliminating the variable $\bar{\mathbf{a}}_i$ from (22) we obtain (20). □

Conversely, it can also be shown that if the convex constraint (20) holds then so does the chance-constraint (5). Below is a sketch of the proof: introduce two variables $\mathbf{a}_i,\bar{\mathbf{a}}_i\ni\mathbf{a}_i+\bar{\mathbf{a}}_i=-y_i\mathbf{w}$ and also let $\bar{a}_{i0}=1-\xi_i+y_ib$. Then LHS of the chance-constraint (5) can be written as:

$$\begin{aligned}\text{LHS of }(5)&=Prob(\mathbf{a}_i^\top X_i+\bar{a}_{i0}+\bar{\mathbf{a}}_i^\top X_i\geq 0)\\ &\leq Prob(\mathbf{a}_i^\top X_i+\bar{a}_{i0}+\max_{\mathbf{x}\in\mathscr{R}_i}\bar{\mathbf{a}}_i^\top\mathbf{x}\geq 0)\\ &=Prob(\mathbf{a}_i^\top X_i+\underbrace{\bar{a}_{i0}+\sum_j\max\left[l_{ij}\bar{a}_{ij},u_{ij}\bar{a}_{ij}\right]}_{a_{io}}\geq 0)\quad =Prob(\mathbf{a}_i^\top X_i+a_{i0}\geq 0)\end{aligned}$$

Now the last probability expression is in the same form as (9). Hence using the arguments in theorem 1 we obtain that if (22) is satisfied, then the original chance-constraint (5) is satisfied. Eliminating $\bar{\mathbf{a}}_i$ from (22) using $\mathbf{a}_i+\bar{\mathbf{a}}_i=-y_i\mathbf{w}$, one obtains (20). Therefore (20) is indeed a valid sufficient condition for the chance-constraint (5) and moreover is a less conservative constraint than (7) by the very construction.

Replacing the chance-constraints (6) in (**CCP**) with the convex constraint (20), we obtain a maximum-margin classification formulation which is robust to uncertain data as well

as less conservative than the **MM-SBMV** formulation. This formulation can be written as the following SOCP:

$$\textbf{(MM-SBMV-I)} \quad \min_{\mathbf{w},b,\xi_i,z_{ij},\tilde{z}_{ij},\mathbf{a}_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{s.t.} \quad 1-\xi_i + y_i b + \sum_j \tilde{z}_{ij} + \sum_j z_{ij} + \kappa\|\Sigma_{(1),i}\mathbf{a}_i\|_2 \leq 0,$$
$$z_{ij} \geq \mu_{ij}^{-} a_{ij},\ z_{ij} \geq \mu_{ij}^{+} a_{ij},\ \xi_i \geq 0,$$
$$\tilde{z}_{ij} \geq -l_{ij}(y_i w_j + a_{ij}), \tilde{z}_{ij} \geq -u_{ij}(y_i w_j + a_{ij}) \quad (23)$$

Note that, the proposed formulations (16,23) are not only robust to the uncertainties in data but are also robust towards moment estimation errors. This is because the formulations employ bounds on mean $(\mu_{ij}^{-}, \mu_{ij}^{+})$ and bounds on second-moment (σ_{ij}^2) rather than the true moments of the uncertain datapoints, which are often unknown.

In the special case where the exact moments of the training datapoints are known, we have that $\mu_i = \mu_i^{-} = \mu_i^{+}$ and $\mathbf{E}[X_{ij}^2] = \sigma_{ij}^2$. Hence the formulation (16) reduces to:

$$\textbf{(MM-SMV)} \quad \min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w}^\top \mu_i - b) \geq 1 - \xi_i + \kappa\|\Sigma_{(2),i}\mathbf{w}\|_2,\ \xi_i \geq 0 \quad (24)$$

where

$$\Sigma_{(2),i} = diag([s_{i1}\nu(\mu_{i1},\mu_{i1},\sigma_{i1}) \ldots s_{in}\nu(\mu_{in},\mu_{in},\sigma_{in})]) \quad (25)$$

Also, in this case, the formulation (23) reduces to:

$$\textbf{(MM-SMV-I)} \quad \min_{\mathbf{w},b,\xi_i,\mathbf{a}_i,\tilde{z}_{ij}} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{s.t.} \quad 1-\xi_i + y_i b + \sum_j \tilde{z}_{ij} + \mu_i^\top \mathbf{a}_i + \kappa\|\Sigma_{(2),i}\mathbf{a}_i\|_2 \leq 0,$$
$$\tilde{z}_{ij} \geq -l_{ij}(y_i w_j + a_{ij}), \tilde{z}_{ij} \geq -u_{ij}(y_i w_j + a_{ij}), \xi_i \geq 0 \quad (26)$$

Note that, the uncertainty sets associated with the formulations (24) and (26) are $\mathscr{E}\left(\mu_i, \kappa\Sigma_{(2),i}\right)$ and $\mathscr{R}_i \cap \mathscr{E}\left(\mu_i, \kappa\Sigma_{(2),i}\right)$ respectively. The subsequent section presents a numerical algorithm for computing the function $\nu(\mu^{-},\mu^{+},\sigma)$ defined in (14).

2.1.1 Computation of $\nu(\mu^{-},\mu^{+},\sigma)$

In this section, we present details of the numerical procedure for computing $\nu(\mu^{-},\mu^{+},\sigma)$ (refer (14)). Recall from lemma 1, the definitions of the normalized random variable and definitions of the corresponding bounds on first $(\hat{\mu}^{\pm})$ and second moment $(\hat{\sigma}^2)$. As noted earlier, we have $|\hat{\mu}^{\pm}| \leq \hat{\sigma} \leq 1$. Now consider the following claim:

Claim Let $\nu(\mu^{-},\mu^{+},\sigma)$ be as defined in (14). Then, $\sqrt{\hat{\sigma}^2 - (\hat{\mu}^{min})^2} \leq \nu(\mu^{-},\mu^{+},\sigma) \leq 1$, where $\hat{\mu}^{min} = \min(|\hat{\mu}^{-}|, |\hat{\mu}^{+}|)$.

Proof Rewriting the definition of $\nu(\mu^-,\mu^+,\sigma)$, we have:

$$\nu(\mu^-,\mu^+,\sigma) = \min\left\{k \geq 0 \;\middle|\; f(t;\hat{\mu},\hat{\sigma},k) \geq 0,\ \forall\, t \in \mathbb{R},\ \forall\, \hat{\mu} \in [\hat{\mu}^-,\hat{\mu}^+]\right\}$$

where $f(t;\hat{\mu},\hat{\sigma},k) \equiv \frac{k^2}{2}t^2 + \max[\hat{\mu}^- t, \hat{\mu}^+ t] - h_{\hat{\mu},\hat{\sigma}}(t)$ (refer (13,12) for definition of $h_{\hat{\mu},\hat{\sigma}}(t)$). Now, let $t \geq 0$ and $f'(t;\hat{\mu}^+,\hat{\sigma},k) = g_1(t) - g_2(t)$ where $g_1(t) \equiv k^2 t + \hat{\mu}^+$ and

$$g_2(t) \equiv \frac{(1-\hat{\mu}^+)\left(\hat{\mu}^+ - \hat{\sigma}^2\right)\exp\left\{t\frac{\hat{\mu}^+ - \hat{\sigma}^2}{1-\hat{\mu}^+}\right\} + \left(\hat{\sigma}^2 - (\hat{\mu}^+)^2\right)\exp\{t\}}{(1-\hat{\mu}^+)^2\exp\left\{t\frac{\hat{\mu}^+ - \hat{\sigma}^2}{1-\hat{\mu}^+}\right\} + (\hat{\sigma}^2 - (\hat{\mu}^+)^2)\exp\{t\}}$$

Now, if $g_1'(0) < g_2'(0)$, then there exists a neighbourhood around $t = 0$ where $f'(t;\hat{\mu}^+,\hat{\sigma}) < 0$ (since $f'(0;\hat{\mu}^+,\hat{\sigma}) = 0$). Also in this neighbourhood $f(t;\hat{\mu}^+,\hat{\sigma}) < 0$ because $f(0;\hat{\mu}^+,\hat{\sigma}) = 0$. Thus $g_1'(0) \geq g_2'(0)$ is a necessary condition for $f \geq 0$. Note that $g_1'(0) = k^2, g_2'(0) = \hat{\sigma}^2 - (\hat{\mu}^+)^2$. Hence, $\nu(\mu^-,\mu^+,\sigma) \geq \sqrt{\hat{\sigma}^2 - (\hat{\mu}^+)^2}$. Similarly, analyzing the case $t \leq 0$ one obtains $\nu(\mu^-,\mu^+,\sigma) \geq \sqrt{\hat{\sigma}^2 - (\hat{\mu}^-)^2}$. Also, from the very definition of $\nu(\mu^-,\mu^+,\sigma)$, we have that its value ≤ 1 (refer lemma 1). This proves the claim. □

Note that, the function $f(t;\hat{\mu},\hat{\sigma},k)$ strictly increases with the value of k and by the above claim we have that $\sqrt{\hat{\sigma}^2 - (\hat{\mu}^{min})^2} \leq k \leq 1$. Thus one can have a simple binary search algorithm for computing $\nu(\mu^-,\mu^+,\sigma)$. The algorithm starts with $k_0^l \equiv \sqrt{\hat{\sigma}^2 - (\hat{\mu}^{min})^2}$ and $k_0^u \equiv 1$. At every iteration, $i \geq 1$, $k_i \equiv \frac{k_{i-1}^l + k_{i-1}^u}{2}$ and it is checked whether

$$f_i^{min} \equiv \left(\min_t f(t;\hat{\mu},\hat{\sigma},k_i)\ \forall\ \hat{\mu} \in [\hat{\mu}^-,\hat{\mu}^+]\right) \geq 0$$

If $f_i^{min} \geq 0$, then $k_i^u \equiv k_i$, else $k_i^l \equiv k_i$. This is repeated until a relevant stopping criteria is met. Checking whether $f_i^{min} \geq 0$ for a fixed value $k_i, \hat{\mu} \in [\hat{\mu}^-,\hat{\mu}^+]$ can be done using any 1-d minimization routine. Also, the criterion is checked at various values of $\hat{\mu} \in [\hat{\mu}^-,\hat{\mu}^+]$. Table 2 shows values of $\nu(\mu^-,\mu^+,\sigma)$ computed using this numerical procedure. For each value of $\hat{\sigma}$, $\nu(\mu^-,\mu^+,\sigma)$ is computed for 10 equally spaced $\hat{\mu}^{\pm}$ values in the range $[-\hat{\sigma},\hat{\sigma}]$. In the table, $\hat{\mu}^-$ and $\hat{\mu}^+$ vary across rows and columns respectively. Hence a '–' represents the case $\hat{\mu}^- > \hat{\mu}^+$ (which is not allowed).

The formulations derived in this section employ partial information of both support and second order moments of uncertainty. These formulations can be specialized to cases where support and mean information alone are available. Though this increases the applicability of the formulations, the resulting classifiers are more conservative as they now employ less information regarding the uncertainties. These specializations are discussed in the subsequent section.

2.2 Formulations using Support and Bounds on Means

In this section, we present a maximum-margin classification formulation which assumes that the bounds on means and the bounding hyper-rectangles (support) for the uncertain training datapoints are known. Though no explicit bounds on second-moments are assumed in this case, the bounding hyper-rectangles imply natural bounds for them: consider the normalized random variable $\hat{X}_{ij} \equiv \frac{X_{ij} - c_{ij}}{s_{ij}}$ studied in lemma 1. It is easy to see that $\mathbf{E}[\hat{X}_{ij}^2] \leq 1$ i.e. $\mathbf{E}[X_{ij}^2] \leq 2\mathbf{E}[X_{ij}]c_{ij} + s_{ij}^2 - c_{ij}^2$. Let us denote this natural bound on $\mathbf{E}[X_{ij}^2]$ as $\left(\sigma_{ij}^*\right)^2$. Now,

Table 2 Values of $v(\mu^-,\mu^+,\sigma)$ computed using the numerical procedure.

$\hat{\sigma}=0.2$									
0.0000	0.3770	0.3875	0.3960	0.3922	0.3796	0.3580	0.3368	0.3233	0.3090
–	0.3770	0.3875	0.3960	0.3922	0.3796	0.3580	0.3367	0.3232	0.3233
–	-1.0000	0.3875	0.3960	0.3922	0.3796	0.3580	0.3368	0.3367	0.3368
–	–	–	0.3960	0.3922	0.3796	0.3582	0.3580	0.3580	0.3580
–	–	–	–	0.3922	0.3796	0.3796	0.3796	0.3796	0.3796
–	–	–	–	–	0.3922	0.3922	0.3922	0.3922	0.3922
–	–	–	–	–	–	0.3960	0.3960	0.3960	0.3960
–	–	–	–	–	–	–	0.3875	0.3875	0.3875
–	–	–	–	–	–	–	–	0.3770	0.3770
–	–	–	–	–	–	–	–	–	0.0000
$\hat{\sigma}=0.4$									
0.0000	0.5096	0.5258	0.5233	0.5191	0.4982	0.4619	0.4088	0.3628	0.3089
–	0.5096	0.5258	0.5233	0.5191	0.4982	0.4619	0.4091	0.3631	0.3628
–	–	0.5258	0.5233	0.5191	0.4982	0.4619	0.4092	0.4091	0.4088
–	–	–	0.5233	0.5191	0.4982	0.4619	0.4619	0.4619	0.4619
–	–	–	–	0.5191	0.4982	0.4982	0.4982	0.4982	0.4982
–	–	–	–	–	0.5191	0.5191	0.5191	0.5191	0.5191
–	–	–	–	–	–	0.5233	0.5233	0.5233	0.5233
–	–	–	–	–	–	–	0.5258	0.5258	0.5258
–	–	–	–	–	–	–	–	0.5096	0.5096
–	–	–	–	–	–	–	–	–	0.0000
$\hat{\sigma}=0.6$									
0.0000	0.6070	0.6656	0.6703	0.6601	0.6362	0.5883	0.5127	0.3911	0.2808
–	0.6070	0.6656	0.6703	0.6601	0.6362	0.5883	0.5127	0.3911	0.3911
–	–	0.6656	0.6703	0.6601	0.6362	0.5883	0.5127	0.5127	0.5127
–	–	–	0.6703	0.6601	0.6362	0.5883	0.5883	0.5883	0.5883
–	–	–	–	0.6601	0.6362	0.6362	0.6362	0.6362	0.6362
–	–	–	–	–	0.6601	0.6601	0.6601	0.6601	0.6601
–	–	–	–	–	–	0.6703	0.6703	0.6703	0.6703
–	–	–	–	–	–	–	0.6656	0.6656	0.6656
–	–	–	–	–	–	–	–	0.6070	0.6070
–	–	–	–	–	–	–	–	–	0.0000
$\hat{\sigma}=0.8$									
0.0000	0.7454	0.7893	0.8224	0.8206	0.7999	0.7543	0.6652	0.5028	0.1524
–	0.7454	0.7893	0.8224	0.8206	0.7999	0.7543	0.6652	0.5028	0.5028
–	–	0.7893	0.8224	0.8206	0.7999	0.7543	0.6652	0.6652	0.6652
–	–	–	0.8224	0.8206	0.7999	0.7543	0.7543	0.7543	0.7543
–	–	–	–	0.8206	0.7999	0.7999	0.7999	0.7999	0.7999
–	–	–	–	–	0.8206	0.8206	0.8206	0.8206	0.8206
–	–	–	–	–	–	0.8224	0.8224	0.8224	0.8224
–	–	–	–	–	–	–	0.7893	0.7893	0.7893
–	–	–	–	–	–	–	–	0.7454	0.7454
–	–	–	–	–	–	–	–	–	0.0000
$\hat{\sigma}=1$									
0.0000	0.8579	0.9418	0.9804	0.9976	0.9938	0.9428	0.8315	0.6285	0.0000
–	0.8579	0.9418	0.9804	0.9976	0.9938	0.9428	0.8315	0.6285	0.6285
–	–	0.9418	0.9804	0.9976	0.9938	0.9428	0.8315	0.8315	0.8315
–	–	–	0.9804	0.9976	0.9938	0.9428	0.9428	0.9428	0.9428
–	–	–	–	0.9976	0.9938	0.9938	0.9938	0.9938	0.9938
–	–	–	–	–	0.9976	0.9976	0.9976	0.9976	0.9976
–	–	–	–	–	–	0.9804	0.9804	0.9804	0.9804
–	–	–	–	–	–	–	0.9418	0.9418	0.9418
–	–	–	–	–	–	–	–	0.8579	0.8579
–	–	–	–	–	–	–	–	–	0.0000

all the formulations presented in the previous section can be specialized using $\sigma_{ij} = \sigma^*_{ij}$. The formulation (16), in this case, reduces to the following SOCP:

$$\boxed{\begin{array}{lll} \textbf{(MM-SBM)} & \min\limits_{\mathbf{w},b,\xi_i,z_{ij}} & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum\limits_{i=1}^{m}\xi_i \\ & \text{s.t.} & 1-\xi_i + y_i b + \sum\limits_j z_{ij} + \kappa\|\Sigma_{(3),i}\mathbf{w}\|_2 \leq 0, \\ & & z_{ij} \geq -y_i\mu_{ij}^- w_j,\, z_{ij} \geq -y_i\mu_{ij}^+ w_j,\; \xi_i \geq 0 \end{array}} \tag{27}$$

where

$$\Sigma_{(3),i} = diag([s_{i1}\nu(\mu_{i1}^-, \mu_{i1}^+, \sigma_{i1}^*)\ldots s_{in}\nu(\mu_{in}^-, \mu_{in}^+, \sigma_{in}^*)]) \tag{28}$$

Also, formulation (23), which employs a less conservative uncertainty set than (16), reduces to:

$$\boxed{\begin{array}{lll} \textbf{(MM-SBM-I)} & \min\limits_{\mathbf{w},b,\xi_i,z_{ij},\tilde{z}_{ij},\mathbf{a}_i} & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum\limits_{i=1}^{m}\xi_i \\ & \text{s.t.} & 1-\xi_i + y_i b + \sum\limits_j \tilde{z}_{ij} + \sum\limits_j z_{ij} + \kappa\|\Sigma_{(3),i}\mathbf{a}_i\|_2 \leq 0, \\ & & z_{ij} \geq \mu_{ij}^- a_{ij},\, z_{ij} \geq \mu_{ij}^+ a_{ij},\; \xi_i \geq 0, \\ & & \tilde{z}_{ij} \geq -l_{ij}(y_i w_j + a_{ij}), \tilde{z}_{ij} \geq -u_{ij}(y_i w_j + a_{ij}) \end{array}} \tag{29}$$

The uncertainty sets associated with the formulations (27, 29) are $\cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(3),i}\right)$ and $\mathscr{R}_i \cap \left(\cup_{\mu_i \in [\mu_i^-, \mu_i^+]} \mathscr{E}\left(\mu_i, \kappa\Sigma_{(3),i}\right)\right)$ respectively.

Interestingly, the value of $\nu(\mu_{ij}^-, \mu_{ij}^+, \sigma_{ij}^*)$ can be computed analytically in the case where $\mu_{ij}^- \leq c_{ij} \leq \mu_{ij}^+$ i.e., the case where the mean, μ_i, is known to lie somewhere around the mid-point of its bounding hyper-rectangle. In particular, if the noise distribution of the i^{th} uncertain datapoint is symmetric, then this assumption is trivially true. The following lemma throws light on this special case:

Lemma 4 *Let the support of a random variable, X, be $[l,u]$ and let the midpoint and semi-length of this interval be denoted by $c \equiv \frac{l+u}{2}$ and $s \equiv \frac{u-l}{2}$ respectively. Let the bounds on the mean $\mathbf{E}[X]$ be $[\mu^-, \mu^+]$and the bound on second-moment be denoted by σ^*. In the case, $\mu^- \leq c \leq \mu^+$, we have:*

$$\nu(\mu^-, \mu^+, \sigma^*) = \sqrt{1 - (\hat{\mu}^{\min})^2}$$

where $\hat{\mu}^{\min} = \min(-\hat{\mu}^-, \hat{\mu}^+)$ and $\hat{\mu}^- \equiv \frac{\mu^- - c}{s}$, $\hat{\mu}^+ \equiv \frac{\mu^+ - c}{s}$.

Proof Recall the definition of the function $h_{\hat{\mu},\hat{\sigma}}(t)$ from (12,13). Here, $\hat{\sigma}$ denotes the upper bound on second-moment of the normalized random variable $\hat{X} = \frac{X-c}{s}$ i.e. $\mathbf{E}[\hat{X}^2] \leq \hat{\sigma}^2$. Since no explicit second-moment bound is assumed in the present case, we have $\hat{\sigma} = 1$. Note that, $h_{\hat{\mu},1}(t) = \log(\cosh t + \hat{\mu}\sinh t),\ t \in \mathbb{R}$. Thus $\nu(\mu^-, \mu^+, \sigma^*)$ (defined in (14)) is the minimum value of k for which $f(t) \geq 0,\ \forall\, t$, where $f(t)$ is defined as follows:

$$f(t) \equiv \begin{cases} \frac{k^2t^2}{2} + \hat{\mu}^+ t - \log\left(\cosh t + \hat{\mu}^+ \sinh t\right),\ t \geq 0 \\ \frac{k^2t^2}{2} + \hat{\mu}^- t - \log\left(\cosh t + \hat{\mu}^- \sinh t\right),\ t < 0 \end{cases}$$

Now, consider the case, $t \geq 0$. Let $f'(t) = g_1(t) - g_2(t)$ where $g_1(t) \equiv k^2 t + \hat{\mu}^+, g_2(t) \equiv \frac{\sinh t + \hat{\mu}^+ \cosh t}{\cosh t + \hat{\mu}^+ \sinh t}$. Now the following claim is true:

Claim $g_2(t)$ is concave for $t \geq 0$.

Proof The value of $g^{''}(t)$ can be calculated as $\frac{8(1-\hat{\mu}^{+2})\exp\{2t\}\left[(1-\hat{\mu}^{+})^2-(1+\hat{\mu}^{+})^2\exp\{4t\}\right]}{[(1+\hat{\mu}^{+})\exp\{2t\}+(1-\hat{\mu}^{+})]^4}$. Also $g^{''}(t) \leq 0 \iff t \geq \frac{1}{2}\log\left\{\frac{1-\hat{\mu}^{+}}{1+\hat{\mu}+}\right\}$. This says that $g_2(t)$ is concave for $t \geq 0$, proving the claim. □

Since $g_2(t)$ is concave, $g_1^{'}(0) \geq g_2^{'}(0)$ implies $f^{'}(t) \geq 0,\ \forall\ t \geq 0$ and is thus a sufficient condition for $f(t) \geq 0\ \forall\ t \geq 0$, as $f(0) = 0$. Also if $g_1^{'}(0) < g_2^{'}(0)$ then $\exists$ a neighbourhood in $t \geq 0$ where $f^{'}(t) < 0$ which implies $f(t) < 0$ in that neighbourhood since $f(0) = 0$. Thus $g_1^{'}(0) \geq g_2^{'}(0)$ is a necessary and sufficient condition for $f(t) \geq 0\ \forall\ t \geq 0$. In other words $k^2 \geq 1-(\hat{\mu}^{+})^2$. Similar arguments for $t < 0$ give the condition $k^2 \geq 1-(\hat{\mu}^{-})^2$. Defining $\hat{\mu}^{\min} = \min(-\hat{\mu}^{-}, \hat{\mu}^{+})$, we have $\nu(\mu^{-}, \mu^{+}, \sigma^{*}) = \sqrt{1-(\mu^{\min})^2}$. This completes the proof. □

Again, in the special case where the means μ_i are known, one has $\mu_i = \mu_i^{-} = \mu_i^{+}$. Using this, the formulation (27) reduces to the following SOCP:

$$\begin{array}{|lll|}\hline \textbf{(MM-SM)} & \min\limits_{\mathbf{w},b,\xi_i} & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum\limits_{i=1}^{m}\xi_i \\ & \text{s.t.} & y_i(\mathbf{w}^{\top}\mu_i - b) \geq 1-\xi_i + \kappa\|\Sigma_{(4),i}\mathbf{w}\|_2,\ \xi_i \geq 0 \\ \hline \end{array} \tag{30}$$

where

$$\Sigma_{(4),i} = diag([s_{i1}\nu(\mu_{i1}, \mu_{i1}, \sigma_{i1}^{*}) \ldots s_{in}\nu(\mu_{in}, \mu_{in}, \sigma_{in}^{*})]) \tag{31}$$

Also, in this case, formulation (29) reduces to:

$$\begin{array}{|lll|}\hline \textbf{(MM-SM-I)} & \min\limits_{\mathbf{w},b,\xi_i,\mathbf{a}_i,\tilde{z}_{ij}} & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum\limits_{i=1}^{m}\xi_i \\ & \text{s.t.} & 1-\xi_i + y_i b + \sum\limits_{j}\tilde{z}_{ij} + \mu_i^{\top}\mathbf{a}_i + \kappa\|\Sigma_{(4),i}\mathbf{a}_i\|_2 \leq 0, \\ & & \tilde{z}_{ij} \geq -l_{ij}(y_i w_j + a_{ij}), \tilde{z}_{ij} \geq -u_{ij}(y_i w_j + a_{ij}), \xi_i \geq 0 \\ \hline \end{array} \tag{32}$$

Note that, the uncertainty sets with the formulations (30) and (32) are $\mathscr{E}\left(\mu_i, \kappa\Sigma_{(4),i}\right)$ and $\mathscr{R}_i \cap \mathscr{E}\left(\mu_i, \kappa\Sigma_{(4),i}\right)$ respectively. It is also interesting to note that in the special cases $\varepsilon = 1$ and $\varepsilon = 0$, the formulation (**MM-SM-I**) degenerates to (**MM-M**) and (**MM-S**) formulations respectively. A comparison of the conservative nature of proposed formulations is presented in the next section.

2.3 Note on the Conservative Nature of the Various Formulations

This section summarizes the formulations presented in the paper and provides a comparison of their conservative nature. The formulations presented in the paper (see Table 1) can be categorized based on whether they employ:

– first order or second order moment information. Formulations named using the symbol '**V**' employ variance (second order moment) information. In general, formulations which employ second order moments are less conservative than those which employ first order moments.

Table 3 Summary of partial information used in various formulations and their corresponding classification formulations and label prediction strategies are indicated.

Form	Representation	Partial Information	Labeling Strategy
1	Single datapoint $\mathbf{d}$	Mean	$y^{pr} = \text{sign}(\mathbf{w}^\top \mathbf{d} - b)$
2	Intervals	Support	$y^{pr} = \text{sign}(\mathbf{w}^\top \mathbf{c}_i - b)$
3	Intervals and moments	Support, Moments	$y^{pr} = \text{sign}(\mathbf{w}^\top \mu_i - b)$
4	Replicates $\mathbf{d}_1, \mathbf{d}_2, \ldots$	Support, Moments	y^{pr} is majority label of replicates Label of replicate: $\text{sign}(\mathbf{w}^\top \mathbf{d}_i - b)$

– bounds or exact moment information. Formulations named using the symbol '**B**' employ bounds on moments rather than the exact moments. In general, formulations which employ moment bounds are more conservative than the ones which exact moments. This is because they also guard against moment estimation errors. However, they are more relevant for real-world data where the exact moments are never known.
– ellipsoidal or intersection of ellipsoidal and hyper-rectangular uncertainty sets. Formulations suffixed with the symbol '**I**' employ uncertainty sets which are intersections of ellipsoidal and hyper-rectangular sets. By construction, these formulations are less conservative than their counterparts, which uses simple ellipsoidal uncertainty sets.

The empirical results in section 4.1 support the comparisons presented here. Hence the proposed formulations and in particular **MM-SMV-I**, lead to robust but not overly-conservative classifiers. Formulations like **MM-SBMV-I** which employ bounds on moments increase the practical applicability of the proposed methodology as the true moments are never known exactly.

3 Classification of Uncertain Test Data

This sections discusses the issue of classifying uncertain test datapoints and presents various error measures for evaluating the performance of classifiers which are robust to uncertain data. As in case of constructing a classifier, different label prediction strategies can be employed based on the level of information available regarding the uncertainty in test datapoints. Table 3 summarizes the various forms in which the uncertainty in datapoints can be represented (here, y^{pr} denotes the predicted label). For each form, the partial information available and the corresponding label prediction methodologies are also indicated. The applicability of a particular classification formulation presented in the paper can be decided based on the partial information available (see Table 1). As noted earlier, even in the cases where either support or moment information is not available at all (e.g. Form 1,2 in table 3), formulation (**MM-SM-I**) can be applied with $\varepsilon = 1$ and $\varepsilon = 0$ respectively. Once a suitable labeling strategy is chosen, the nominal error, **NomErr**, can be calculated as percentage of wrongly classified test datapoints:

$$\textbf{NomErr} = \frac{\sum_i 1_{y_i^{pr} \neq y_i}}{\#\ test\ datapoints} \times 100 \tag{33}$$

Note that, since each test datapoint is uncertain, there is always some (non-zero) probability that it is misclassified unless the entire bounding hyper-rectangle lies on the correct side of the discriminating hyperplane. This misclassification probability is clearly not considered by the nominal error. Based on the discussion presented in section 2.1, error measures which take into account the per test datapoint misclassification probabilities can be

Table 4 Expressions of ε_{opt} for different error measures $\mathbf{OptErr}_i^{1-4,c}$

Error Measure	Formulation	Uncertainty Region	ε_{opt}
$\mathbf{OptErr_i^1}$	MM-SBMV	$\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)$	$\exp\left\{-\frac{(\mathbf{w}^\top\mu_i^{opt}-b)^2}{2\left(\mathbf{w}^\top\Sigma^2_{(1),i}\mathbf{w}\right)}\right\}$
$\mathbf{OptErr_i^2}$	MM-SMV	$\mathscr{E}\left(\mu_i,\kappa\Sigma_{(2),i}\right)$	$\exp\left\{-\frac{(\mathbf{w}^\top\mu_i-b)^2}{2\left(\mathbf{w}^\top\Sigma^2_{(2),i}\mathbf{w}\right)}\right\}$
$\mathbf{OptErr_i^3}$	MM-SBM	$\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(3),i}\right)$	$\exp\left\{-\frac{(\mathbf{w}^\top\mu_i^{opt}-b)^2}{2\left(\mathbf{w}^\top\Sigma^2_{(3),i}\mathbf{w}\right)}\right\}$
$\mathbf{OptErr_i^4}$	MM-SM	$\mathscr{E}\left(\mu_i,\kappa\Sigma_{(4),i}\right)$	$\exp\left\{-\frac{(\mathbf{w}^\top\mu_i-b)^2}{2\left(\mathbf{w}^\top\Sigma^2_{(4),i}\mathbf{w}\right)}\right\}$
$\mathbf{OptErr_i^c}$	MM-MC	$\mathscr{E}\left(\mu_i,\kappa_c\Sigma_i^{\frac{1}{2}}\right)$	$\frac{\mathbf{w}^\top\Sigma_i\mathbf{w}}{\left(\mathbf{w}^\top\mu_i-b\right)^2+\mathbf{w}^\top\Sigma_i\mathbf{w}}$

derived. To this end, consider an uncertain test datapoint X_i with label y_i is given and let $\mu_{ij}^{opt}=\arg\max_{\mu\in[\mu_{ij}^-,\mu_{ij}^+]}(-y_i\mu w_j)$. Then by Theorem 1, the true probability of misclassification of the test datapoint X_i will be less than or equal to ε (i.e. $Prob\left[y_i(\mathbf{w}^\top X_i-b)\leq 0\right]\leq\varepsilon$) if:

$$y_i(\mathbf{w}^\top\mu_i^{opt}-b)\geq\kappa\|\Sigma_{(1),i}\mathbf{w}\|_2$$

The above inequality is arrived at by setting $\xi_i=1$ in (7) and re-arranging terms. Using this, one can calculate the least value of $\varepsilon=\varepsilon_{opt}$ for which the above inequality is satisfied:

$$\varepsilon_{opt}=\exp\left\{-\frac{(\mathbf{w}^\top\mu_i^{opt}-b)^2}{2\left(\mathbf{w}^\top\Sigma^2_{(1),i}\mathbf{w}\right)}\right\}$$

By lemma 2, ε_{opt} is the value of ε for which the uncertainty region $\cup_{\mu_i\in[\mu_i^-,\mu_i^+]}\mathscr{E}\left(\mu_i,\kappa\Sigma_{(1),i}\right)$ touches the discriminating hyperplane, $\mathbf{w}^\top\mathbf{x}-b=0$. Also, by the very definition of ε_{opt}, the true probability of misclassification of the test datapoint X_i will be less than or equal to it. This leads to the following error definition on each test datapoint:

$$\mathbf{OptErr}_i=\begin{cases}1 & \text{if } y_i\neq y_i^{pr}\\ \varepsilon_{opt} & \text{if } y_i=y_i^{pr} \text{ and } \exists\,\mathbf{x}\in\mathscr{R}_i\ni y_i(\mathbf{w}^\top\mathbf{x}-b)<0\\ 0 & \text{if } y_i(\mathbf{w}^\top\mathbf{x}-b)\geq 0\;\forall\,\mathbf{x}\in\mathscr{R}_i\end{cases}\tag{34}$$

Now the analysis in the previous paragraph can be repeated with various uncertainty regions considered in this paper. The error measures with different uncertainty regions differ only in computing the value of ε_{opt}. The uncertainty regions, the corresponding error measures and ε_{opt} are summarized in Table 4. An error measure derived using the Chebyshev based relaxation (denoted by $\mathbf{OptErr}_i^{c5}$) is also presented in the table. The overall error, $\mathbf{OptErr}^{1\text{-}4,c}$, can be calculated as percentage of $\mathbf{OptErr}_i^{1\text{-}4,c}$ over all test datapoints:

$$\mathbf{OptErr}^{1\text{-}4,c}=\frac{\sum_i\mathbf{OptErr}_i^{1\text{-}4,c}}{\#\ test\ datapoints}\times 100\tag{35}$$

Note that, both **NomErr** and $\mathbf{OptErr}^{1\text{-}4,c}$ can be estimated for any hyperplane classifier, provided the partial information employed is available. Also, since the $\mathbf{OptErr}^{1\text{-}4,c}$

[5] Here Σ_i denotes the covariance matrix of the i^{th} uncertain datapoint

error measures employ the Bernstein/Chebyshev probability bounds, they represent the upper bound on the true probability of misclassification averaged over the test datapoints. Since probability bounds are employed, the absolute values of these errors will invariably be higher than the corresponding **NomErr** values. However since the **NomErr** error measure completely neglects the probability of misclassification arising due to the uncertainty in test datapoints, the new error measures $\mathbf{OptErr}^{1\text{-}4,c}$ provide a "finer" estimate of the true testset error. Experimental results reported in the subsequent section show that the proposed classifiers achieve lower **NomErr** and $\mathbf{OptErr}^{1\text{-}4,c}$ when compared to existing uncertain data classifiers.

4 Numerical experiments

This section presents experimental results which extensively compare the proposed (see Table 1) and existing methodologies (see section 1.1) for classifying uncertain data. The following datasets were used in the experiments:

WBCD Wisconsin Breast Cancer Diagnostic dataset[6]. The task is to classify "benign" and "malignant" tumours based on 10 features computed from tumour cell nuclei. However, since the measurements are not the same over all tumour cells, the mean, standard-error and maximum values of the 10 features are provided. From this information the support and moments for each training datapoint are estimated. Bounds on the means, $[\mu^-, \mu^+]$, are estimated using the standard-error information. This dataset is an example of Form-3 data (see Table 3).

Micro-array Task is to identify four kinds of drugs: Azoles ($\mathscr{A}$), Fibrates ($\mathscr{F}$), Statins ($\mathscr{S}$) and Toxicants ($\mathscr{T}$) based on gene-expression data[7] [13]. Since the experiments are noisy, three replicates of each datapoint are provided. Instead of handling a multi-class problem we have defined six binary classification tasks using "one versus one" scheme (e.g. $\mathscr{A}$ vs. $\mathscr{F}$ and so on). As a preprocessing step, we have reduced the dimension of the problem to 166 by feature selection using Fisher score.

Synthetic Generation methodology: a) nominal (true) datapoints were generated using Gaussian mixture models b) uncertainty was introduced into each nominal point using standard finite-supported distributions (whose parameters were chosen randomly) c) replicates for each nominal datapoint were produced by sampling the chosen noise distribution. The synthetic datasets are named using dimension of the dataset and are subscripted with the distribution used for generating replicates (e.g. synthetic data of dimensionality n with Uniform, truncated Beta, skew-Normal and skew-t noise distributions are denoted by $\mathbf{n_U}$, $\mathbf{n}_\beta$, $\mathbf{n_{SN}}$ and $\mathbf{n_{ST}}$ respectively).

Both the **Micro-array** and **Synthetic** datasets stand as examples of Form-4 data (refer table 3). Also, in these cases, the support and moments for each datapoint were estimated from the corresponding replicates. The bounds on first moments ($[\mu^-, \mu^+]$) and bounds on second-moments (σ^2) were estimated using the Hotelling's T^2-statistic (see e.g. page 227, [10]) and Cochran's theorem (see e.g. page 419, [17]) respectively.

As the key motivation is to develop robust as well as non-overly-conservative classifiers, the first set of experiments, presented in section 4.1, compare the conservative nature of various robust classification constraints derived in the paper and existing in the literature. In

[6] Available at `http://mlearn.ics.uci.edu/MLSummary.html`

[7] Available at `http://www.ncbi.nlm.nih.gov/geo/` with accession number GSE2187.

Table 5 List of various relaxations of the chance-constraint

Relaxation Scheme	**Formulation**	Relaxed Constraint
Bernstein bounding	**MM-SMV** (24)	$b \leq \mathbf{w}^\top \mu_i - \kappa \|\Sigma_{(2),i}\mathbf{w}\|_2$
Bernstein bounding	**MM-SMV-I** (26)	$b \leq \max(\mathbf{w}^\top \mu_i - \kappa \|\Sigma_{(2),i}\mathbf{w}\|_2, \mathbf{w}^\top \mathbf{c}_i - \|\mathbf{S}_i\mathbf{w}\|_1)$
Bernstein bounding	**MM-SM** (30)	$b \leq \mathbf{w}^\top \mu_i - \kappa \|\Sigma_{(4),i}\mathbf{w}\|_2$
Bernstein bounding	**MM-SM-I** (32)	$b \leq \max(\mathbf{w}^\top \mu_i - \kappa \|\Sigma_{(4),i}\mathbf{w}\|_2, \mathbf{w}^\top \mathbf{c}_i - \|\mathbf{S}_i\mathbf{w}\|_1)$
Chebyshev bounding	**MM-MC** (4)	$b \leq \mathbf{w}^\top \mu_i - \kappa_c \|\Sigma_i^{\frac{1}{2}}\mathbf{w}\|_2$
Mean based	**MM-M** (2)	$b \leq \mathbf{w}^\top \mu_i$
Support based	**MM-S** (3)	$b \leq \mathbf{w}^\top \mathbf{c}_i - \|\mathbf{S}_i\mathbf{w}\|_1$

particular, these experiments compare the conservative nature of Chebyshev and Bernstein bounding schemes for relaxing CCP based learning formulations. As the results show, Bernstein schemes lead to far less conservative relaxations than Chebyshev schemes and hence have potential to be exploited in building robust classifiers for uncertain data. Section 4.2 presents experiments which compare the margin, $2/\|\mathbf{w}\|_2$, achieved by the proposed and existing robust classifiers on synthetic datasets. The results show that the proposed classifiers achieve higher margin and hence have the potential to generalize better.

Section 4.3 presents the key empirical results of the paper — comparison of various robust classifiers discussed in this paper using the error measures **NomErr** (33) and **OptErr**$^{\mathbf{1\text{-}4,c}}$ (35). Results show that in case of all datasets, the proposed classifiers achieve better generalization than state-of-the-art.

As mentioned earlier, classifiers derived using Bernstein relaxation schemes are also inherently robust to moment estimation errors. This is because the proposed classifiers require knowledge of moment bounds rather than the exact moments themselves. Section 4.4 presents experiments comparing the robustness of various uncertain data classifiers to moment estimation errors. The results show that the proposed classifiers are less susceptible to moment estimation errors than existing classifiers.

4.1 Comparison of the Conservative Nature of the Various Robust Constraints

In this section we compare the conservative nature of the various robust classification constraints presented in this paper. In particular, we compare conservativeness of the various convex relaxations of the chance-constraint $Prob(\mathbf{w}^\top X_i - b \leq 0) \leq \varepsilon$. Note that this constraint is a variant of the original chance-constraint (5) with $y_i = 1$ and the $1 - \xi_i$ term neglected. Table 5 summarizes various relaxations of this chance-constraint derived using the Bernstein and Chebyshev bounding schemes. Also, the constraints (2) in (**MM-M**) and (3) in (**MM-S**) are accordingly modified and are shown in the table. These represent the two extreme relaxations — most lenient and most conservative. The relaxations which use bounds on moments are not compared here in order to have a fair comparison with Chebyshev schemes — which can only be employed if exact moments are known. Constraints shown in the table for **MM-SMV-I** and **MM-SM-I** can be derived easily from Lemma 3, rather than from constraints in the corresponding formulations (26,32).

Now, suppose the value of $\mathbf{w}$ is fixed. Then, the conservative nature of the various relaxations can be compared by looking at the least upper bound on b. Greater the value of the least upper bound on b, lesser is the conservativeness of the corresponding relaxation. Noting this observation, the following experiment was done: in each run of the experiment a random vector $\mathbf{w}$ was chosen and datapoints were sampled from a random distribution.

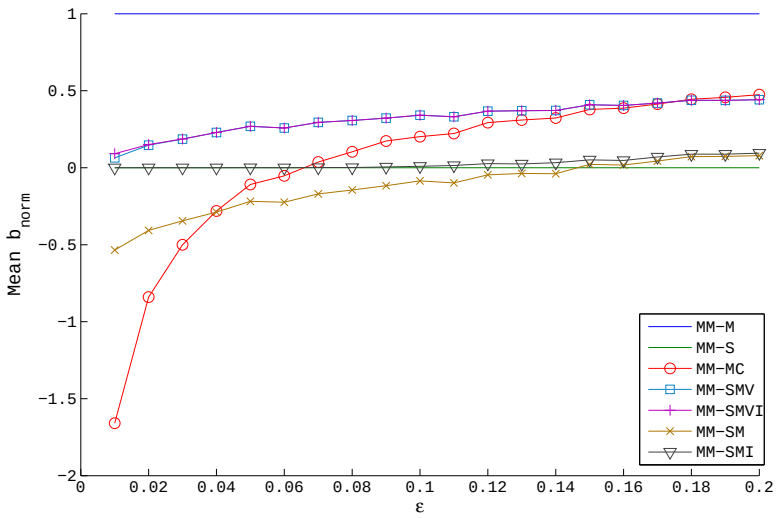

Fig. 1 A plot of mean b_{norm} vs. ε comparing different robust constraints.

Random distributions were simulated by employing random parameters for the truncated skewed t-distribution. Using the sampled datapoints, support and second order moments were estimated. Employing this partial information and the chosen value of $\mathbf{w}$, the least upper bound on b with various constraints can be calculated using table 5[8]. Now let the value of b thus obtained with **MM-M** and **MM-S** be b_m and b_s respectively. Figure 1 shows the plot of $b_{norm} \equiv \frac{b-b_s}{b_m-b_s}$ averaged over 50 such experimental runs at different values of ε. Since ε is a small number denoting upper bound on misclassification probability, only values of $\varepsilon \in [0, 0.2]$ are interesting and hence are shown in the figure. Note that the value of b_{norm} with **MM-SMV-I** and **MM-SMV** is the highest — proving that the Bernstein relaxation schemes are less conservative than the Chebyshev based schemes. It is interesting to note that for $\varepsilon \leq 0.04$, even the Bernstein relaxations using first order moments ((30) in **MM-SM** and (32) in **MM-SM-I**) are less conservative than the second order moment based Chebyshev relaxations ((4) in **MM-MC**). Hence, formulations derived using the proposed methodology model uncertainty in a less conservative fashion and are expected to achieve better generalization while being robust to uncertainties in data.

4.2 Comparison of Classification Margin

This section presents experimental results which compare various robust formulations based on the classification margin ($2/\|\mathbf{w}\|_2$) achieved by them. Figure 2 shows plots of margin vs. ε with various formulations at a fixed value of $C = 10$ on different synthetic datasets. Clearly, the margins achieved by the proposed formulations at all ε values are higher than those achieved with **MM-MC** and **MM-S**. Since higher margins imply better generalization, the proposed classifiers are expected to generalize better. Also note that for higher dimensional datasets the gain in margin with the proposed classifiers is higher. The plots also show that the proposed formulations model uncertainty in a less conservative fashion, regardless

[8] Recall that $\mathbf{c}_i$ represents the geometric center of $\mathscr{R}_i$ and $\mathbf{S}_i$ is the diagonal matrix with entries as semi-lengths of the bounding hyper-rectangle $\mathscr{R}_i$

of the underlying noise distribution. This is expected as the proposed methodology does not make any distributional assumptions. The margin achieved by **MM-M** is highest as it neglects the uncertainty of the datapoints and assumes mean as the only possible position for the datapoint. Also, the trends shown in Figure 2 remained the same at different values of the C parameter.

4.3 Comparison of Generalization Error

This section presents the experimental results comparing $\mathbf{OptErr^{1\text{-}4,c}}$ (35) and **NomErr** (33) incurred by the various robust classifiers presented in this paper. In cases where the uncertainty in datasets is represented using replicates (e.g. **Micro-array** data), the comparison is also done with an SVM constructed assuming each replicate as a training datapoint. We denote this classifier as **MM-R**. The results are summarized in table 6. In each case, the hyper-parameters (C and/or ε) were tuned using a 5-fold cross-validation procedure. The reported error values represent the cross-validation error obtained using the corresponding tuned set of hyper-parameters averaged over three 5-fold cross-validation experiments. Hence lower the values of $\mathbf{OptErr^{1\text{-}4,c}}$ and **NomErr**, better is the generalization ability of the corresponding robust classifier. Clearly, the proposed classifiers incurred the least error on all the datasets. Moreover in terms of the $\mathbf{OptErr^{1\text{-}4,c}}$ measures, the proposed classifiers outperformed state-of-the-art in case of most of the datasets.

The results also show that the nearest competitor to the proposed classifiers is the **MM-MC** classifier — which is also based on chance-constrained techniques. The results hence show that, in general, formulations based on chance-constrained techniques and moreover the ones using the Bernstein bounds are best suited for handling uncertainties in data. As mentioned earlier, the classifiers derived using the Bernstein relaxations require knowledge of bounds on moments rather than the exact moments. Clearly, the results in table 6 show that the variants which employ bounds on moments achieve better generalization than the ones which employ exact moments. This inherent advantage of the proposed methodology is again illustrated in the subsequent section.

4.4 Robustness to Moment Estimation Errors

In this section, we present experiments comparing various formulations for robustness to moment estimation errors. The experimental set-up is as follows: a) A synthetic dataset template (e.g. $\mathbf{n_{ST}}$) is chosen b) Using the same template 10 different training sets are generated. Hence the training sets differ only in terms of the replicates; nominal datapoints are the same. c) Independently a testset consisting of nominal datapoints alone is also generated. Hence the testset is not noisy and represents the true data. d) Various classifiers are trained (with fixed values of hyper-parameters) using each of the 10 training sets, and **NomErr** on the testsets was noted. The standard deviation in the testset error incurred on few synthetic data templates is summarized in table 7. Ideally, since each of the 10 training sets represent the same "true" (nominal) set of datapoints, the variation in the testset accuracy must be zero. The results show the variation in testset error is least for **MM-SBMV-I** and **MM-SBM-I** — which are the variants employing bounds on moments. This proves that the proposed classifiers are robust to moment estimation errors and illustrates the benefit of the proposed methodology. It is also interesting to note that, the variation in testset accuracies with **MM-SBMV-I** and **MM-SBM-I** is less than that with **MM-R** and **MM-S**.

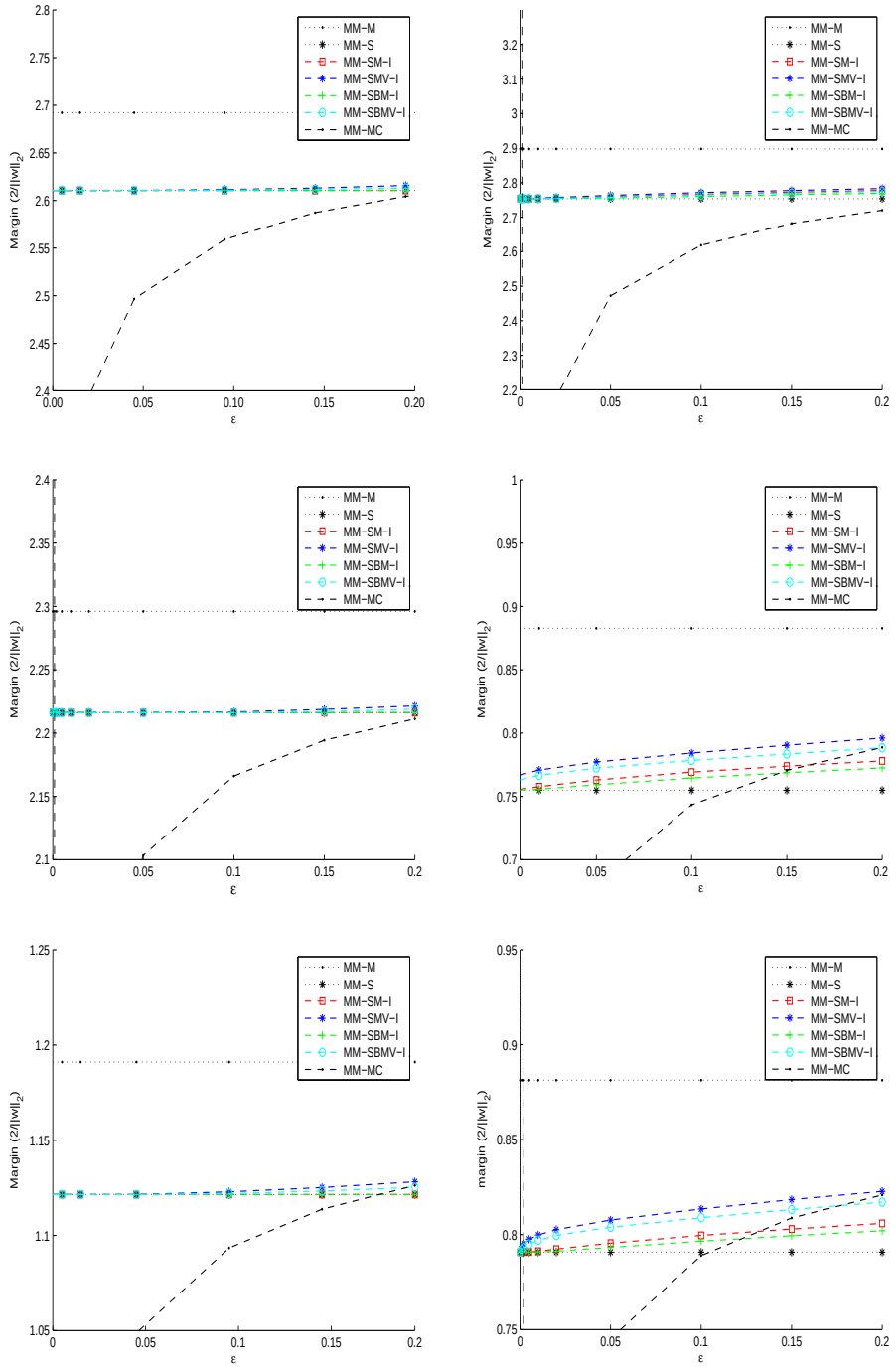

Fig. 2 Figure comparing margins achieved by 8 formulation at various ε values (2B on top-left, 10B on top-right,2U on middle-left, 10U on middle-right, 2SN on bottom-left, 10SN on bottom-right)

.

Table 6 Comparison of **OptErr**$^{1\text{-}4,c}$ and **NomErr** for various robust classifiers.

Dataset:	**WBCD**							
	MM-M	**MM-R**	**MM-S**	**MM-MC**	**MM-SM-I**	**MM-SMV-I**	**MM-SBM-I**	**MM-SBMV-I**
OptErr^{1}	37.26	–	40.50	37.26	37.26	37.26	**29.20**	29.55
OptErr^{2}	37.26	–	45.82	37.26	37.26	37.26	**32.40**	32.98
OptErr^{3}	37.26	–	45.02	37.26	37.26	37.26	**32.20**	32.72
OptErr^{4}	37.26	–	45.82	37.26	37.26	37.26	32.60	**32.41**
OptErr^{c}	37.26	–	40.70	37.26	37.26	37.26	**28.82**	29.18
NomErr	55.67	–	**37.26**	55.67	**37.26**	**37.26**	**37.26**	**37.26**
Dataset:	$\mathscr{A}$ vs. $\mathscr{F}$							
	MM-M	**MM-R**	**MM-S**	**MM-MC**	**MM-SM-I**	**MM-SMV-I**	**MM-SBM-I**	**MM-SBMV-I**
OptErr^{1}	41.57	40.98	58.34	36.36	36.38	36.38	34.48	**34.41**
OptErr^{2}	39.21	38.55	58.32	34.23	34.25	34.25	**32.37**	32.50
OptErr^{3}	46.13	45.60	58.38	40.28	40.39	40.39	**37.91**	38.07
OptErr^{4}	46.00	45.46	58.38	40.17	40.26	40.26	**37.80**	37.96
OptErr^{c}	77.18	77.97	**58.97**	71.5	**58.97**	**58.97**	**58.97**	**58.97**
NomErr	**00.00**	**00.00**	55.29	**00.00**	**00.00**	**00.00**	**00.00**	**00.00**
Dataset:	$\mathscr{A}$ vs. $\mathscr{S}$							
	MM-M	**MM-R**	**MM-S**	**MM-MC**	**MM-SM-I**	**MM-SMV-I**	**MM-SBM-I**	**MM-SBMV-I**
OptErr^{1}	61.10	65.13	61.69	53.26	53.42	53.42	51.73	**51.29**
OptErr^{2}	59.25	63.35	61.69	51.88	52.02	52.02	50.04	**49.66**
OptErr^{3}	64.74	68.65	61.69	56.21	56.41	56.41	55.09	**54.61**
OptErr^{4}	64.64	68.56	61.69	56.13	56.32	56.32	54.99	**54.52**
OptErr^{c}	86.24	88.66	**61.69**	80.64	**61.69**	**61.69**	**61.69**	**61.69**
NomErr	09.02	08.65	61.69	06.10	06.10	**5.71**	06.10	06.52
Dataset:	$\mathscr{A}$ vs. $\mathscr{T}$							
	MM-M	**MM-R**	**MM-S**	**MM-MC**	**MM-SM-I**	**MM-SMV-I**	**MM-SBM-I**	**MM-SBMV-I**
OptErr^{1}	67.33	75.00	58.33	54.37	53.95	53.95	**47.22**	50.74
OptErr^{2}	65.54	73.32	58.33	52.97	52.68	52.68	**46.77**	50.08
OptErr^{3}	70.61	78.03	58.33	56.89	56.37	56.37	**47.22**	51.67
OptErr^{4}	70.51	77.94	58.33	56.81	56.28	56.28	**47.22**	51.67
OptErr^{c}	88.94	93.23	58.33	79.33	58.33	58.33	**47.22**	51.67
NomErr	09.72	**06.67**	58.33	08.33	08.33	09.72	08.89	11.11
Dataset:	$\mathscr{F}$ vs. $\mathscr{S}$							
	MM-M	**MM-R**	**MM-S**	**MM-MC**	**MM-SM-I**	**MM-SMV-I**	**MM-SBM-I**	**MM-SBMV-I**
OptErr^{1}	31.23	35.21	47.68	24.52	24.76	24.76	**22.03**	22.15
OptErr^{2}	29.27	33.32	46.89	22.71	22.96	22.96	**20.24**	20.38
OptErr^{3}	34.99	38.86	49.31	28.17	28.39	28.39	**25.70**	25.79
OptErr^{4}	34.86	38.73	49.25	28.05	28.27	28.27	**25.57**	25.66
OptErr^{c}	69.91	73.70	**62.95**	66.51	**62.95**	**62.95**	**62.95**	**62.95**
NomErr	01.03	00.95	28.21	**00.00**	**00.00**	**00.00**	00.99	00.99
Dataset:	$\mathscr{F}$ vs. $\mathscr{T}$							
	MM-M	**MM-R**	**MM-S**	**MM-MC**	**MM-SM-I**	**MM-SMV-I**	**MM-SBM-I**	**MM-SBMV-I**
OptErr^{1}	51.69	54.70	63.89	41.08	41.45	41.45	**36.54**	36.88
OptErr^{2}	49.84	52.81	63.87	39.63	40.07	40.07	**35.48**	35.79
OptErr^{3}	55.14	58.37	63.89	43.76	44.03	44.03	**38.51**	38.88
OptErr^{4}	55.02	58.25	63.89	43.66	43.94	43.94	**38.43**	38.80
OptErr^{c}	79.84	83.52	63.89	71.16	63.89	63.89	**60.34**	60.78
NomErr	06.55	**03.81**	58.17	05.63	05.28	05.28	06.75	06.11
Dataset:	$\mathscr{S}$ vs. $\mathscr{T}$							
	MM-M	**MM-R**	**MM-S**	**MM-MC**	**MM-SM-I**	**MM-SMV-I**	**MM-SBM-I**	**MM-SBMV-I**
OptErr^{1}	61.85	68.00	69.35	42.15	43.13	43.13	**39.41**	39.54
OptErr^{2}	60.31	66.25	69.35	41.03	41.98	41.98	**38.28**	38.41
OptErr^{3}	64.71	71.18	69.35	44.22	45.30	45.30	**41.41**	41.56
OptErr^{4}	64.63	71.09	69.35	44.14	45.21	45.21	**41.33**	41.48
OptErr^{c}	85.48	90.39	69.35	70.40	69.35	69.35	**67.73**	67.81
NomErr	08.28	06.28	69.35	05.95	**05.63**	06.30	05.97	06.97

Table 7 Standard deviation in testset error, **NomErr**, incurred by various robust classifiers.

	MM-M	MM-R	MM-S	MM-MC	MM-SM-I	MM-SMV-I	MM-SBM-I	MM-SBMV-I
$\mathbf{10_{ST}}$	0.7160	0.8551	0.6779	0.7738	0.5996	0.5731	**0.5389**	0.5446
$\mathbf{15_{ST}}$	0.4908	0.4698	0.1829	0.4118	0.2222	0.2833	0.2042	**0.1504**
$\mathbf{20_{ST}}$	0.8553	0.4517	0.2396	0.5864	0.2081	0.4286	**0.1853**	0.2086

5 Conclusions

A novel methodology for constructing robust classifiers by employing partial information on the support and moments of the uncertain training datapoints was presented. The idea was to pose the uncertain data classification problem as a CCP and relax it as a convex SOCP formulation using Bernstein bounding schemes. The key advantage of the Bernstein relaxation scheme is to model uncertainty in a less conservative manner. Moreover, since the relaxation requires the knowledge of bounds on moments rather than the exact moments themselves, the resulting classifiers are also inherently robust to moment estimation errors. Using the proposed methodology, various robust formulations employing different levels of partial information were derived. Interesting error measures for evaluating performance of classifiers robust to uncertain data were also presented. The performance of the proposed classifiers was empirically evaluated on various synthetic and real-world datasets.

The main conclusions to be drawn from the experimental results are as follows: 1) In general, the Bernstein relaxation schemes are less conservative than the Chebyshev based schemes. This key feature was exploited in the proposed methodology for developing classifiers that model data uncertainty very efficiently. In future, it would be interesting to explore the applicability of Bernstein schemes for relaxing various other CCP-based learning formulations. 2) Classifiers developed using the proposed methodology achieve higher margins and hence better generalization than state-of-the-art. 3) Formulation using bounds on moments (**MM-SBM-I, MM-SBMV-I**) not only achieve good generalization but are less susceptible to moment estimation errors.

Acknowledgements The authors CB and SB would like to acknowledge grants from Yahoo! and IBM Research Labs.

References

1. Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics, 2009. In Press.
2. Aharon Ben-Tal and Arkadi Nemirovski. Selected Topics in Robust Convex Optimization. *Mathematical Programming*, 112(1):125–158, 2007.
3. 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, 2004.
4. J. Bi and T. Zhang. Support Vector Classification with Input Data Uncertainty. In *Advances in Neural Information Processing Systems*, 2004.
5. W. Chen and M. Sim. Goal Driven Optimization. To appear in Operations Research.
6. X. Chen, M. Sim, and P. Sun. A Robust Optimization Perspective on Stochastic Programming. *Operations Research*, 55(6):1058–1071.
7. X. Chen, M. Sim, P. Sun, and CP. Teo. From CVaR to Uncertainty Set: Implications in Joint Chance Constrained Optimization. To appear in Operations Research.
8. Francesca Demichelis, Paolo Magni, Paolo Piergiorgi, Mark A Rubin, and Riccardo Bellazzi. A Hierarchical Nave Bayes Model for Handling Sample Heterogeneity in Classification Problems: An Application to Tissue Microarrays. *BMC Bioinformatics*, 7(514), 2006.

9. L. E. Ghaoui, G. R. G. Lanckriet, and G. Natsoulis. Robust Classification with Interval Data. Technical Report UCB/CSD-03-1279, Computer Science Division, University of California, Berkeley, 2003.
10. Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall; 5th edition, 2002.
11. G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A Robust Minimax Approach to Classification. *Journal of Machine Learning Research*, 3:555–582, 2003.
12. J. Saketha Nath, C. Bhattacharyya, and M. N. Murty. Clustering based Large Margin Classification: A Scalable Approach using SOCP Formulation. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 674–679, New York, NY, USA, 2006. ACM Press.
13. Georges Natsoulis, Laurent El Ghaoui, Gert R.G. Lanckriet, Alexander M. Tolley, Fabrice Leroy, Shane Dunlea, Barrett P. Eynon, Cecelia I. Pearson, Stuart Tugendreich, and Kurt Jarnagin. Classification of a Large Microarray Data Set: Algorithm Comparison and Analysis of Drug Signatures. *Genome Research*, 15:724–736, 2005.
14. Arkadi Nemirovski and Alexander Shapiro. Convex Approximations of Chance Constrained Programs. *SIAM Journal of Optimization*, 17(4):969–996, 2006.
15. Y. Nesterov and A. Nemirovskii. *Interior Point Polynomial Algorithms in Convex Programming*. Number 13. Studies in Applied and Numerical Mathematics, SIAM books, PA 19104-2688, USA, 1993.
16. R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
17. Henry Scheffé. *The Analysis of Variance*. Wiley-IEEE, 1959.
18. Pannagadatta K. Shivaswamy, Chiranjib Bhattacharyya, and Alexander J. Smola. Second Order Cone Programming Approaches for Handling Missing and Uncertain Data. *Journal of Machine Learning Research*, 7:1283–1314, 2006.
19. V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.