CORC Technical Report TR-2004-10
Ambiguous chance constrained problems and robust optimization[*]

E. Erdoğan[†] G. Iyengar[‡]

September 7, 2004

Abstract

In this paper we study ambiguous chance constrained problems where the distributions of the random parameters in the problem are themselves uncertain. We primarily focus on the special case where the uncertainty set $\mathcal{Q}$ of the distributions is of the form $\mathcal{Q} = \{\mathbb{Q} : \rho_p(\mathbb{Q}, \mathbb{Q}_0) \leq \beta\}$, where ρ_p denotes the Prohorov metric. The ambiguous chance constrained problem is approximated by a robust sampled problem where each constraint is a robust constraint centered at a sample drawn according to the central measure $\mathbb{Q}_0$. The main contribution of this paper is to show that the robust sampled problem is a good approximation for the ambiguous chance constrained problem with high probability. This result is established using the Strassen-Dudley Representation Theorem that states that when the *distributions* of two random variables are close in the Prohorov metric one can construct a *coupling* of the random variables such that the *samples* are close with high probability. We also show that the robust sampled problem can be solved efficiently both in theory and in practice.

1 Introduction

A large class of decision problems in finance and engineering can be formulated as optimization problems of the form

$$\begin{array}{rl} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & f(\mathbf{x}, \mathbf{h}) = \max_{1 \leq i \leq l} \{f_i(\mathbf{x}, \mathbf{h})\} \leq 0, \\ & \mathbf{x} \in \mathcal{X}, \end{array} \tag{1}$$

where $\mathbf{x} \in \mathcal{X}$ is the decision vector, $\mathcal{X} \subseteq \mathbf{R}^n$ is a closed convex set, $\mathbf{h} \in \mathbf{R}^m$ are problem parameters and each of the functions $f_i(\mathbf{x}, \mathbf{h}) : \mathcal{X} \times \mathcal{H} \mapsto \mathbf{R}$ are convex in $\mathbf{x}$ for a fixed $\mathbf{h}$. We assume, without loss of generality, that the objective is linear and independent of $\mathbf{h}$.

[*] Submitted to *Math. Prog.*

[†] IEOR Department, Columbia University, New York, New York 10027. Email: ee168@columbia.edu. Research partially supported by NSF grant CCR-00-09972.

[‡] IEOR Department, Columbia University, New York, New York 10027. Email: garud@columbia.edu. Research partially supported by NSF grants CCR-00-09972, DMS-01-04282 and ONR grant N000140310514.

The deterministic optimization approach to solving optimization problems computes a solution to (1) assuming that the problem parameters $\mathbf{h}$ are known and fixed. In practice, however, the parameters $\mathbf{h}$ are the result of some measurement or estimation process, and are, therefore, never known for certain. This uncertainty is of serious concern in applications because solutions of optimization problems are often very sensitive to fluctuations in the problem parameters. This phenomenon is well documented in several different application areas [3, 26].

Recently *robust optimization* has emerged as an attractive optimization framework for reducing the sensitivity of the optimal solution to perturbations in the parameter values. In this framework, the uncertain parameters $\mathbf{h}$ are assumed to belong to a bounded uncertainty set $\mathcal{H}$ and the *robust optimization problem* corresponding to the *nominal problem* (1) is given by

$$\begin{array}{ll} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & f(\mathbf{x},\mathbf{h}) \leq 0, \quad \forall\, \mathbf{h} \in \mathcal{H}, \\ & \mathbf{x} \in \mathcal{X} \end{array} \tag{2}$$

This framework was introduced in Ben-Tal and Nemirovski [4, 5, 6]. There is also a parallel literature on robust formulations of optimization problems originating from robust control [18, 20]. In many applications the uncertainty set $\mathcal{H}$ is given by the confidence region around the point estimates of the parameter allowing one to provide probabilistic guarantees on the performance of the optimal solution of the robust problem [26]. The robust problem (2) is solved by reformulating the semi-infinite constraints, $f(\mathbf{x},\mathbf{h}) \leq 0$, for all $\mathbf{h} \in \mathcal{H}$, as a finite collection of constraints. Such a reformulation is only possible when the uncertainty set $\mathcal{H}$ and the function $f(\mathbf{x},\mathbf{h})$ satisfy some regularity conditions. See [4, 6, 7] for robust formulations that can be solved efficiently. Even when the reformulation is possible, the resulting problem is typically harder than the nominal problem (1) ([7] proposes a new framework wherein the robust problem remains in the same complexity class as the nominal problem). In general, however, the robust problem is intractable.

Another criticism of the robust framework is that it gives the same "weight" to all perturbations $\mathbf{h} \in \mathcal{H}$. Also, in certain applications one might have the flexibility of violating the constraints corresponding to a small fraction of the set $\mathcal{H}$. An alternative optimization framework that mitigates this criticism to some extent is called *chance-constrained optimization.* In this framework, one assumes the parameters are $\mathbf{h}$ are distributed according to a known distribution $\mathbb{Q}$ on $\mathcal{H}$, and replaces the nominal problem (1) by the following *chance-constrained problem*

$$\begin{array}{ll} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & \mathbf{x} \in \mathcal{X}_\epsilon(\mathbb{Q}), \end{array} \tag{3}$$

where

$$\mathcal{X}_\epsilon(\mathbb{Q}) = \Big\{ \mathbf{x} \in \mathcal{X} : \mathbb{Q}(\mathbf{H} : f(\mathbf{x},\mathbf{H}) > 0) \leq \epsilon \Big\}, \tag{4}$$

for some $0 < \epsilon < 1$. The parameter ϵ controls the probability that the optimal solution of (3) violates the constraints – as $\epsilon \downarrow 0$ the chance-constrained problem starts to resemble the robust problem (2). Although chance-constrained problems have a long history dating back to at least the work of Charnes and Cooper [12], they have not found wide applicability. This is primarily because computing the optimal solution for chance-constrained problems is extremely hard. To begin with just evaluating $\mathbb{Q}(\mathbf{H} : f(\mathbf{x}, \mathbf{H}) > 0)$ involves a multidimensional integral that becomes hard as the number of parameters grows. Moreover, even if the function $f(\mathbf{x}, \mathbf{h})$ is convex (or even linear) in $\mathbf{x}$ the feasible set $\mathcal{X}_\epsilon(\mathbb{Q})$ of (3) is not convex. A detailed discussion of the chance-constrained programs and, more generally, stochastic programs can be found in [35].

Recently, Calafiore and Campi [10, 11] and de Farias and Van Roy [14] independently proposed tractable approximations to (3) based on constraint sampling and statistical learning techniques. In this approach, one approximates chance-constrained problem (3) by the following *sampled problem*

$$\begin{array}{rl} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & f(\mathbf{x}, \mathbf{H}_i) \leq 0, \quad i = 1, ..., N, \\ & \mathbf{x} \in \mathcal{X}, \end{array} \tag{5}$$

where $\mathbf{H}_i$, $i = 1, \ldots, N$, are N independent, identically distributed (IID) samples from the distribution $\mathbb{Q}$. de Farias and Van Roy [14] consider the special case where $f(\mathbf{x}, \mathbf{h}) = \mathbf{h}^T\mathbf{x} + h_0$ and use results from Computational Learning Theory [1, 30, 43] to show that for all $N \geq \frac{4n}{\epsilon} \ln\left(\frac{12}{\epsilon}\right) + \frac{4}{\epsilon} \ln\left(\frac{2}{\delta}\right)$, the feasible set of the sampled problem (5) is contained in $\mathcal{X}_\epsilon(\mathbb{Q})$ with probability at least $1 - \delta$. Thus, in this sampling based method there are two possible sources of errors: with probability δ, the feasible set of (5) (and consequently, the optimal solution of (5)) may not be contained in $\mathcal{X}_\epsilon(\mathbb{Q})$; and, in event that this is not the case, the feasible points of (5) can still violate the constraint $f(\mathbf{x}, \mathbf{H}) \leq 0$ with a probability ϵ. The analysis in [14] can be extended to general $f(\mathbf{x}, \mathbf{h})$ (see Section 3 for details). Calafiore and Campi [10, 11] consider general convex functions $f(\mathbf{x}, \mathbf{h})$ and show that for $N \geq \frac{2n}{\epsilon} \ln\left(\frac{12}{\epsilon}\right) + \frac{2}{\epsilon} \ln\left(\frac{2}{\delta}\right) + 2n$, the optimal solution of the sampled problem (5) is feasible for (3) with probability at least $1 - \delta$. On the one hand, this bound is weak in the sense that it is only valid for the optimal solution, and *not* the entire feasible set. On the other hand, the number of samples required to ensure that the optimal solution is feasible for (3) with high probability can be orders of magnitude lower. The result in [10, 11] is proved using a fundamental fact that the optimal solution of a convex program is "supported" by at most n constraints. We will briefly review this work in Section 3.3. Recently, Nemirovski and Shapiro [33, 32] established logarithmically separated upper and lower bounds on the number of samples required to approximate a chance constrained problem when the measure $\mathbb{Q}$ has well defined moment generating function.

Although the bounds on the sample size N are distribution-free in the sense that they do not depend on the underlying measure $\mathbb{Q}$, in order to construct the sampled problem (5) one requires samples from this probability measure. Also, there is an implicit assumption that the distribution $\mathbb{Q}$ of the random parameters $\mathbf{H}$ is fixed. A major criticism raised against chance constrained problems and, more generally, stochastic programs is that, in practice, the measure is never known exactly. Just as the point estimates for the parameters, the distribution $\mathbb{Q}$ is also estimated from data or measurements, and is, therefore, known only to within some error, i.e. the measure $\mathbb{Q} \in \mathcal{Q}$ where $\mathcal{Q}$ is a set of measures. Since our primary interest in the chance constrained problem (3) was to use it as an approximation (or even a refinement) of the robust problem (2), the natural problem to consider when the measure $\mathbb{Q}$ is uncertain is given by

$$\begin{array}{ll} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & \mathbf{x} \in \bar{\mathcal{X}}_\epsilon, \end{array} \tag{6}$$

where

$$\bar{\mathcal{X}}_\epsilon = \Big\{ \mathbf{x} \in \mathcal{X} : \mathbb{Q}(\mathbf{H} : f(\mathbf{x}, \mathbf{H}) > 0) \leq \epsilon, \ \forall \mathbb{Q} \in \mathcal{Q} \Big\}. \tag{7}$$

We will call (6) an *ambiguous chance-constrained problem.* A problem of the form (6) has two sources of uncertainty: the distribution $\mathbb{Q}$ of the parameter $\mathbf{h}$ is uncertain, and, given a measure $\mathbb{Q}$, the particular realization of the parameter $\mathbf{h}$ is also uncertain. In the decision theory literature the uncertainty in the distribution is referred to as *ambiguity*, and hence the name for the problem.

Modeling ambiguity and its consequence has been receiving attention in several different fields. The minimax formulation has a long history in stochastic programing [44, 8, 16, 17, 29, 40, 38, 39]. Ruszczynski and Shapiro [36] show the equivalence between minimax stochastic programming and minimizing a convex risk measure [2, 23] of the second-stage cost. [37] extends the minimax approach to a multiperiod setting. The study of ambiguity in Economics began with the work of Gilboa and Schmeidler [25]. This work was extended to a multiperiod setting by Hansen and Sargent [27] and Epstein and his co-authors [13, 21, 22]. Ambiguity in the context of Markov decision processes was independently investigated by Iyengar [28] and El Ghaoui and Nilim [19].

The main contributions of this paper are as follows.

(a) We consider uncertainty sets $\mathcal{Q}$ of measures that are of the form $\mathcal{Q} = \{\mathbb{Q} : \rho(\mathbb{Q}, \mathbb{Q}_0) \leq \beta\}$ where $\rho(\cdot,\cdot)$ denotes a suitable metric between probability measures, i.e. the uncertainty sets are "balls" centered around the central measure $\mathbb{Q}_0$. We approximate the ambiguous chance-constrained problem (6) by a *robust sampled problem* defined as follows

$$\begin{array}{ll} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & f(\mathbf{x}, \mathbf{z}) \leq 0, \quad \forall\ \mathbf{z} \text{ s.t. } \|\mathbf{z} - \mathbf{H}_i^0\| \leq \beta, \quad i = 1, \ldots, N, \end{array} \tag{8}$$

where $\mathbf{H}_i^0$, $i = 1, \ldots, N$, denote IID samples drawn according to the central measure $\mathbb{Q}_0$ and the norm $\|\cdot\|$ on the $\mathcal{H}$ space is related to the probability metric $\rho(\cdot,\cdot)$ (details are given in Section 4). Results in [7] imply that for a large class of constraint functions $f(\mathbf{x}, \mathbf{h})$ and suitably defined norms $\|\cdot\|$ the robust sampled problem (8) is in the same complexity class as the nominal problem (1).

(b) We combine results from Computational Learning Theory with results for *coupling* of random variables [42] to compute upper bounds on the number of samples N required to ensure that the feasible set of the robust sampled problem (8) is contained in $\bar{\mathcal{X}}_\epsilon$ with high probability. This bound depends on the Vapnik-Chervonenkis (VC) dimension of the function $f(\mathbf{x}, \mathbf{h})$.

(c) We use coupling to extend the results of Calafiore and Campi [10, 11] to the ambiguous chance constrained problems, i.e. we compute upper bounds on the number of samples required to ensure that the optimal solution of the robust sampled problem (8) is contained in $\bar{\mathcal{X}}_\epsilon$ with high probability. The bound in this case depends on the number of "support" constraints, and is independent of the VC dimension of $f(\mathbf{x}, \mathbf{h})$.

The issue of ambiguity of measures was also raised in [11] where the authors considered a finite uncertainty set $\mathcal{Q}$. They proposed a solution strategy where one samples from *all* of these measures and showed that the samples from different measure "help" each other. In contrast, we consider the case where $\mathcal{Q}$ is uncountably infinite and we draw samples from *only* the central measure $\mathbb{Q}_0$.

The rest of this paper is structured as follows. In Section 3 we briefly review the known results for chance constrained problem. The results in this section are not new – they have been included to set the context for our extensions. Section 4 introduces probability metrics, coupling and the Strassen-Dudley Representation Theorem. Section 5 uses this Representation Theorem to establish bounds for ambiguous chance constrained problems. In Section 6 we identify particular classes of functions $f(\cdot,\cdot)$ and norms $\|\cdot\|$ on the parameter space $\mathcal{H}$ that allow the robust sampled problem (8) to be solved efficiently. Section 7 has some concluding remarks.

2 Notation

Sets will be denoted by calligraphic letters, e.g. $\mathcal{A}$. For a finite set $\mathcal{A}$, we will denote the size of $\mathcal{A}$ by $|\mathcal{A}|$. All (deterministic) vectors will be denoted by boldface lowercase letters, e.g. $\mathbf{x}$. Random vectors and samples of random vectors will be denoted by boldface uppercase letters, e.g. $\mathbf{H}$, and measures will be denoted by mathematical boldface letters, e.g. $\mathbb{P}$. We will denote that a random vector $\mathbf{H}$ has distribution $\mathbb{Q}$ by $\mathbf{H} \sim \mathbb{Q}$, a σ-algebra on a space $\mathcal{H}$ by $\mathcal{F}(\mathcal{H})$, and the set of all probability measures on $\mathcal{H}$ by $\mathcal{M}(\mathcal{H})$. We will denote the n-th binomial coefficient $\frac{N!}{(N-n)!n!}$ by $\binom{N}{n}$.

3 Chance constrained problems and Learning Theory

In this section our goal is to relate the sampled problem (5) to the chance constrained problem (3). We assume that the distribution $\mathbb{Q}$ of the perturbations $\mathbf{H}$ is known and fixed. Let $\mathbf{H}_{1,N} = \{\mathbf{H}_1, \mathbf{H}_2, ..., \mathbf{H}_N\}$ denote N IID samples of the random vector $\mathbf{H} \sim \mathbb{Q}$. Then the feasible set of the sampled problem (5) is given by

$$\mathcal{Y}[\mathbf{H}_{1,N}] = \{\mathbf{x} \in \mathcal{X} : f(\mathbf{x}, \mathbf{H}_i) \leq 0, i = 1, ..., N\} \tag{9}$$

In the sequel, we will often abbreviate $\mathcal{Y}[\mathbf{H}_{1,N}]$ as $\mathcal{Y}[N]$ with the understanding that the set $\mathcal{Y}[N]$ is defined using a *particular* sequence of IID samples $\mathbf{H}_{1,N}$ of length N. In the first half of this section we appropriately interpret concepts from Computational Learning Theory [1, 43, 30] to establish bounds on the number of samples N required for $\mathcal{Y}[N]$ to be a good approximation for the feasible set $\mathcal{X}_\epsilon = \{\mathbf{x} : \mathbb{Q}(\mathbf{H} : f(\mathbf{x},\mathbf{H}) \leq 0) \leq \epsilon\}$ of the chance constrained problem (3). Next, we briefly summarize the result in [10, 11] that provides a bound for the number of samples required to ensure that the optimal solution of the sampled problem (5) is contained in $\mathcal{X}_\epsilon$ with high probability. As mentioned in Section 1, the results in this section are not new – they simply provide the context for the new results in Section 5.

3.1 Decision vectors, concepts and Vapnik-Chervonenkis (VC) dimension

With each decision vector $\mathbf{x} \in \mathcal{X}$ we associate the *concept* or *classification*

$$\mathcal{C}_x = \{\mathbf{h} \in \mathcal{H} : f(\mathbf{x},\mathbf{h}) \leq 0\}.$$

Let $\mathcal{C}_f = \{\mathcal{C}_x : \mathbf{x} \in \mathcal{X}\}$ denote the class of all concepts induced on $\mathcal{H}$ as the decision vector $\mathbf{x}$ runs over the set $\mathcal{X}$. Thus, the set $\mathcal{X}_\epsilon = \{\mathbf{x} \in \mathcal{X} : \mathbb{Q}(\mathbf{H} \not\in \mathcal{C}_x) \leq \epsilon\}$. To apply the results from Learning Theory to the problem at hand we pretend that our goal is to learn the set $\mathcal{X}_0$ of concepts $\mathcal{C}_x$ that cover $\mathcal{H}$ with probability 1, i.e. $\mathcal{X}_0 = \{\mathbf{x} : \mathbb{Q}(\mathbf{H} \in \mathcal{C}_x) = 1\}$. Since Learning algorithms only have access to a finite number of samples of the random variable $\mathbf{H}$, it is impossible to learn the concepts in $\mathcal{X}_0$; instead any such algorithm will have to be satisfied with learning a concept with a small error ϵ, i.e. a concept $\mathcal{C}_x$ with $\mathbb{Q}(\mathbf{H} \not\in \mathcal{C}_x) \leq \epsilon$ or equivalently $\mathbf{x} \in \mathcal{X}_\epsilon$. For the particular case considered in this paper, learning $\mathcal{X}_\epsilon$ is equivalent to producing a good approximation for the function $f(\mathbf{x},\cdot)$ using a finite number of samples. Thus, one should expect that the complexity of learning $\mathcal{X}_\epsilon$ when the function $f(\mathbf{x},\mathbf{h}) = \mathbf{h}^T\mathbf{x}$ is linear, or equivalently the associated concept $\mathcal{C}_x$ is a half space, should be smaller than the complexity of learning $\mathcal{X}_\epsilon$ when the function $f(\cdot,\cdot)$ is nonlinear. Learning Theory quantifies the complexity of a concept class $\mathcal{C}_f$ by its *Vapnik-Chervonenkis* (VC) dimension [43].

Let $\mathcal{S} = \{\mathbf{s}_1,\ldots,\mathbf{s}_N\} \subset \mathcal{H}$ denote a finite subset of $\mathcal{H}$ with $|S| = N$. Define

$$\Pi_f(\mathcal{S}) = \left\{ \left(\mathbf{1}_{\mathcal{C}_x}(\mathbf{s}_1),\ldots,\mathbf{1}_{\mathcal{C}_x}(\mathbf{s}_N)\right) : \mathbf{x} \in \mathcal{X} \right\}, \tag{10}$$

where $\mathbf{1}_{\mathcal{C}}(\cdot)$ denotes the characteristic function of the set $\mathcal{C}$. The set Π_f is the set of *dichotomies* or *behaviors* induced by the concept class $\mathcal{C}_f$, or equivalently the function $f(\cdot,\cdot)$. From (10), it is clear that the number of elements $|\Pi_f(\mathcal{S})| \leq 2^N$. We say that a set $\mathcal{S}$ is *shattered* by the concept class $\mathcal{C}_f$ if $\Pi_f(\mathcal{S}) = \{0,1\}^N$, or equivalently $|\Pi_f(\mathcal{S})| = 2^N$. Note that if a set $\mathcal{S}$ is shattered by the concept class $\mathcal{C}_f$ it does not yield any information about the concept class. Thus, the size of largest shattered set is a measure of the complexity of the concept class $\mathcal{C}_f$.

Definition 1 (VC dimension of $f(\cdot,\cdot)$) *The VC dimension d_f of the function $f(\cdot,\cdot)$ is the cardinality of the largest set $\mathcal{S} \subset \mathcal{H}$ that is shattered by the concept class $\mathcal{C}_f$, i.e*

$$\begin{aligned} d_f &= \sup\left\{|\mathcal{S}| : \Pi_f(\mathcal{S}) = \{0,1\}^N\right\}, \\ &= \sup\left\{|\mathcal{S}| : |\Pi_f(\mathcal{S})| = 2^N\right\}. \end{aligned} \tag{11}$$

In the sequel we will find it convenient to work with the growth function $\pi_f(N)$ defined as follows.

$$\pi_f(N) = \max\left\{|\Pi_f(\mathcal{S})| : |\mathcal{S}| = N\right\}. \tag{12}$$

The growth function π_f is another measure of the complexity of the concept class: the faster this function grows, the more behaviors on sets of size m that can be realized by $\mathcal{C}_f$; consequently, the less is the information that this finite set conveys about the class $\mathcal{C}_f$. A surprising and fundamental result in Computational Learning Theory asserts that if the VC dimension $d_f < \infty$, the growth function $\pi_f(N)$ is bounded by a polynomial in N of degree d_f.

Proposition 1 (Sauer's Lemma [9, 1, 30]) *Suppose the VC dimension d_f of the function $f(\cdot,\cdot)$ is finite. Then*

$$\pi_f(N) \leq 1 + \binom{N}{1} + \binom{N}{2} + \ldots + \binom{N}{d_f} \leq \left(\frac{eN}{d_f}\right)^{d_f}, \tag{13}$$

where e denotes the base of natural logarithm.

In this paper we assume that the VC dimension $d_f < \infty$. This is not a very restrictive assumption since many functions $f(\cdot,\cdot)$ used in practice have finite VC dimension.

Proposition 2 *Let d_f denote the VC dimension of the function $f(\cdot,\cdot)$.*

(a) $\mathcal{X} = \mathbf{R}^n$, $\mathcal{H} = \{(h_0, \mathbf{h}) : h_0 \in \mathbf{R}, \mathbf{h} \in \mathbf{R}^n\} = \mathbf{R}^{n+1}$ and $f(\mathbf{x}, \mathbf{h}) = h_0 + \mathbf{h}^T\mathbf{x}$. Then $d_f \leq n$.

(b) $\mathcal{X} = \mathbf{R}^n$, $\mathcal{H} = \{(\mathbf{A}, \mathbf{b}, \mathbf{u}, v) : \mathbf{A} \in \mathbf{R}^{p\times n}, \mathbf{b}, \mathbf{c} \in \mathbf{R}^n, v \in \mathbf{R}\}$, and $f(\mathbf{x}, \mathbf{h}) = \|\mathbf{A}\mathbf{x} + \mathbf{b}\| - \mathbf{u}^T\mathbf{x} - v$. Then $d_f \leq O(n^2)$.

(c) Suppose the VC dimension of the function $f_i(\cdot,\cdot)$ is d_i, $i = 1, \ldots, l$. Then the VC dimension d_f of the function $f(\mathbf{x}, \mathbf{h}) = \max_{1\leq i\leq l}\{f_i(\mathbf{x}, \mathbf{h})\}$ is bounded above by $d_f \leq \mathcal{O}(10^l \max_{1\leq i\leq l}\{d_i\})$.

Proof: Part (a) is proved on p.77 in [1] (see also [14]), part (b) is established in [9] and part (c) can be established using techniques in [31]. ■

Part (c) states that the best known bound on the VC dimension of a pointwise maximum of l functions grows *exponentially* in l. Thus, the VC dimension of the concept class induced by constraint function $f(\cdot,\cdot)$ of the nominal problem (1) can be quite large. We will remark on this at the end of the next section.

3.2 Learning the chance constrained set $\mathcal{X}_\epsilon$

For $\mathbf{x} \in \mathcal{X}$ let $\text{err}(\mathbf{x}) = \mathbb{Q}(\mathbf{H} \notin \mathcal{C}_x)$. Thus, $\mathcal{X}_\epsilon = \{\mathbf{x} \in \mathcal{X} : \text{err}(\mathbf{x}) \leq \epsilon\}$. The feasible set $\mathcal{Y}[N]$ of the sampled problem (5) is the set of all decision vectors $\mathbf{x}$, or equivalently concepts $\mathcal{C}_x$, that are consistent with the given sample $\mathbf{H}_{1,N}$. Intuitively speaking, if the sample size is large enough one would expect that $\mathcal{Y}[N]$ is a *good* estimate of the set $\mathcal{X}_\epsilon$. The next two results make this rigorous.

Lemma 1 *Fix $\epsilon > 0$. Suppose $\bar{\mathbf{x}} \in \mathcal{X}$ with err$(\bar{\mathbf{x}}) > \epsilon$. Then, for all $N \geq 1$,*

$$\mathbb{Q}^N \left(\mathbf{H}_{1,N} : \bar{\mathbf{x}} \in \mathcal{Y}[N] \right) \leq e^{-\epsilon N}, \tag{14}$$

where $\mathbb{Q}^N$ denotes the product measure $\mathbb{Q} \times \mathbb{Q} \times \ldots \times \mathbb{Q}$ with N terms.

Proof: Recall that $\mathbf{H}_{1,N}$ are IID samples of the random vector $\mathbf{H} \sim \mathbb{Q}$. Therefore,

$$\mathbb{Q}^N \left(\mathbf{H}_{1,N} : \bar{\mathbf{x}} \in \mathcal{Y}[N] \right) = \left(\mathbb{Q}(\mathbf{H} : f(\bar{\mathbf{x}}, \mathbf{H}) \leq 0) \right)^N \leq (1 - \epsilon)^N \leq e^{-\epsilon N},$$

where the last inequality follows from the fact that $1 - \epsilon \leq e^{-\epsilon}$. ■

Lemma 1 establishes that the probability that a given concept $\mathcal{C}_x$ with err$(\mathbf{x}) > \epsilon$ is contained in $\mathcal{Y}[N]$ decays exponentially with the number of samples N. Suppose the set $\mathcal{X}$ is finite. Then the union bound implies that $\mathbb{Q}^N \left(\mathbf{H}_{1,N} : \mathcal{Y}[N] \not\subseteq \mathcal{X}_\epsilon \right) \leq |X| e^{-\epsilon N} \leq \delta$, for all $N \geq \frac{1}{\epsilon} \log \left(\frac{|X|}{\delta} \right)$, i.e $\mathcal{O}\left(\frac{1}{\epsilon} \log \left(\frac{|X|}{\delta} \right) \right)$ samples are needed to learn $\mathcal{X}_\epsilon$ with a probability of error bounded by δ. Since the complexity of learning a concept is determined by the VC dimension of the concept class, we expect that a similar bound should hold with $|X|$ replaced by $\pi_f(N)$.

Lemma 2 (Proposition 8.2.3 in [1]) *Let π_f denote the growth function associated with concept class $\mathcal{C}_f$ induced by $f(\cdot, \cdot)$. Then, for all $N \geq 8/\epsilon$,*

$$\mathbb{Q}^N \left(\mathbf{H}_{1,N} : \mathcal{Y}[N] \not\subseteq \mathcal{X}_\epsilon \right) \leq 2\pi_f(2N) 2^{-\epsilon N/2}. \tag{15}$$

This result and the upper bound (13) imply the following corollary.

Corollary 1 *Fix $\epsilon, \delta > 0$. Suppose the VC dimension d_f of $f(\cdot, \cdot)$ is finite. Then*

$$\mathbb{Q}^N \left(\mathbf{H}_{1,N} : \mathcal{Y}[N] \not\subseteq \mathcal{X}_\epsilon \right) \leq \delta,$$

for all

$$N \geq \max \left\{ \frac{8}{\epsilon}, \left(\frac{4d_f}{\epsilon} \log \left(\frac{12}{\epsilon} \right) + \frac{4}{\epsilon} \log \left(\frac{2}{\delta} \right) \right) \right\}.$$

We conclude this section with the following lower bound.

Lemma 3 (Theorem 3.5 in [30]) *Suppose the VC dimension d_f of the function by $f(\cdot, \cdot)$ is finite. Then the worst case number of samples required to learn $\mathcal{X}_\epsilon$ is $\Omega(d_f/\epsilon)$.*

Corollary 1 and Lemma 3 establish that the number of samples $N = \Theta(d_f/\epsilon)$. From Proposition 2 (c) we have that the VC dimension of the constraint $f(\cdot, \cdot)$ in the nominal problem (1) could be as large as $10^l \max_{1 \leq i \leq l}\{d_i\}$ where d_i is VC dimension of the functions f_i, $i = 1, \ldots, l$. Thus, the number of samples required to learn $\mathcal{X}_\epsilon$ could be prohibitive even for well behaved constraint functions.

3.3 Quality of the optimal solution of the sampled problem

In this section the goal is more modest – we want to compute the number of samples required to ensure that only the *optimal* solution of the sampled problem (5), as opposed to the entire set $\mathcal{Y}[N]$, is feasible for the chance constrained problem (3) with high probability. Calafiore and Campi [10, 11] recently showed that $N = \mathcal{O}(n/\epsilon)$ is enough to achieve this goal. In this section we briefly review the results in [10, 11].

Let (P) denote the following convex program

$$\begin{array}{rl} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, N, \end{array}$$

where $f_i(\mathbf{x})$ is a convex function of $\mathbf{x}$ for all $i = 1, \ldots, N$. Let $\widehat{\mathbf{x}}$ denote the *unique* optimal solution of (P). Let (P_k) denote the convex program obtained by dropping the k-th constraint, $k = 1, \ldots, N$, and let $\widehat{\mathbf{x}}_k$ denote the *unique* optimal solution of the problem (P_k). See [10, 11] for the case where the optimal solutions are not unique.

Definition 2 (Support constraint) *The k-th constraint $f_k(\mathbf{x}) \leq 0$ is called a support constraint for the problem (P) if $\mathbf{c}^T\widehat{\mathbf{x}}_k < \mathbf{c}^T\widehat{\mathbf{x}}$.*

Theorem 3 (Theorem 2 in [11]) *The convex program has at most n support constraints.*

Lemma 4 *Fix $\epsilon > 0$. Let $\widehat{\mathbf{x}}$ denote the optimal solution of the sampled problem (5). Then $\mathbb{Q}^N(\mathbf{H}_{1,N} : \widehat{\mathbf{x}} \notin \mathcal{X}_\epsilon) \leq \binom{N}{n} e^{-\epsilon(N-n)}$.*

Proof: The sampled problem (5) is a convex program with N constraints. Let $\mathcal{I} \subseteq \{1, \ldots, N\}$ with $|I| = n$. Let $\mathcal{H}_\mathcal{I}^N = \{(\mathbf{h}_1, \ldots, \mathbf{h}_N) : (\mathbf{h}_i)_{i\in\mathcal{I}} \text{ are the support constraints}\}$. Then Theorem 3 implies $\mathcal{H}^N = \cup_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|=n\}} \mathcal{H}_\mathcal{I}^N$. Thus,

$$\begin{aligned} \mathbb{Q}^N(\mathbf{H}_{1,N} : \widehat{\mathbf{x}} \notin \mathcal{X}_\epsilon) &= \sum_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|=n\}} \mathbb{Q}^N(\mathbf{H}_{1,N} \in \mathcal{H}_\mathcal{I}^N : \widehat{\mathbf{x}}_\mathcal{I} \notin \mathcal{X}_\epsilon) \\ &= \sum_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|=n\}} \Big(\mathbb{Q}^n(\mathbf{H}_{i\in I} : \widehat{\mathbf{x}}_\mathcal{I} \notin \mathcal{X}_\epsilon) \prod_{i\notin\mathcal{I}} \mathbb{Q}(\mathbf{H}_i : f(\widehat{\mathbf{x}}_\mathcal{I}, \mathbf{H}_i) \leq 0|\mathcal{A}_\mathcal{I})\Big), \end{aligned}$$

where $\widehat{\mathbf{x}}_\mathcal{I}$ denotes the optimal solution of the sampled problem (5) with only the samples $i \in \mathcal{I}$ present, $\mathcal{A}_\mathcal{I}$ is the event $\mathcal{A}_\mathcal{I} = \{\mathbf{H}_{i\in I} : \widehat{\mathbf{x}}_\mathcal{I} \notin \mathcal{X}_\epsilon\}$ and each probability in the sum can be written as a product because $\mathbf{H}_{1,N}$ are IID samples. Since $\widehat{\mathbf{x}} \notin \mathcal{X}_\epsilon$, it follows that $\mathbb{Q}(\mathbf{H}_i : f(\widehat{\mathbf{x}}_\mathcal{I}, \mathbf{H}_i) \leq 0|\mathcal{A}_\mathcal{I}) \leq (1-\epsilon)$, for all $i \notin \mathcal{I}$. Thus,

$$\begin{aligned} \mathbb{Q}^N(\mathbf{H}_{1,N} : \widehat{\mathbf{x}} \notin \mathcal{X}_\epsilon) &\leq (1-\epsilon)^{(N-n)} \sum_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|=n\}} \mathbb{Q}^n(\mathbf{H}_{i\in I} : \widehat{\mathbf{x}}_\mathcal{I} \notin \mathcal{X}_\epsilon) \\ &\leq \binom{N}{n}(1-\epsilon)^{(N-n)} \quad \leq \quad \binom{N}{n} e^{-\epsilon(N-n)}. \end{aligned}$$

■

Lemma 4 immediately implies the following.

Corollary 2 *Fix $\epsilon, \delta > 0$. Let $\widehat{\mathbf{x}}$ denote the optimal solution of the sampled problem (5). Then*

$$\mathbb{Q}^N\left(\mathbf{H}_{1,N} : \widehat{\mathbf{x}} \not\subseteq \mathcal{X}_\epsilon\right) \leq \delta,$$

for all

$$N \geq \frac{2n}{\epsilon}\log\left(\frac{1}{\epsilon}\right) + \frac{2}{\epsilon}\log\left(\frac{1}{\delta}\right) + 2n$$

4 Probability metrics and Coupling

In Section 1 we had introduced the following robust chance constrained set (see (7))

$$\bar{\mathcal{X}}_\epsilon = \Big\{\mathbf{x} \in \mathcal{X} : \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{Q}(\mathbf{H} : f(\mathbf{x}, \mathbf{H}) > 0) \leq \epsilon\Big\},$$

where $\mathcal{Q} = \{\mathbb{Q} : \rho(\mathbb{Q}, \mathbb{Q}_0) \leq \beta\}$ for an appropriately chosen metric ρ on the space $\mathcal{M}(\mathcal{H})$ of probability measures on $\mathcal{H}$. Recall that $\mathcal{X} \subseteq \mathbf{R}^n$, $\mathcal{H} \subseteq \mathbf{R}^m$ and we denote the norm in $\mathcal{H}$ space by $\|\cdot\|$. In this section we first review properties of some basic probability metrics. Next, we introduce the concept of *coupling* of random variables that plays an important role in constructing approximations of the robust chance constrained set $\bar{\mathcal{X}}_\epsilon$ via samples. In this paper we will be primarily using the Prohorov metric ρ_p defined as follows.

$$\rho_p(\mathbb{Q}^1, \mathbb{Q}^2) = \inf\left\{\epsilon : \mathbb{Q}^1(\mathcal{B}) \leq \mathbb{Q}^2(\mathcal{B}^\epsilon) + \epsilon, \forall \mathcal{B} \in \mathcal{F}(\mathcal{H})\right\}, \tag{16}$$

where

$$\mathcal{B}^\epsilon = \Big\{\mathbf{x} \in \mathcal{X} : \inf_{\mathbf{z} \in \mathcal{B}} \|\mathbf{x} - \mathbf{z}\| \leq \epsilon\Big\}.$$

Although the definition appears asymmetric, ρ_p is a metric. It plays an important role in probability because it metrizes weak convergence. Moreover, $\rho_p(\mathbb{Q}_1, \mathbb{Q}_2)$ is the minimum distance "in probability" between random variables distributed according to $\mathbb{Q}_i$, $i = 1, 2$. Some other metrics of interest are as follows.

(a) Wassestein or Kantorovich metric ρ_w:

$$\rho_w(\mathbb{Q}_1, \mathbb{Q}_2) = \sup\left\{\left|\int_{\mathcal{H}} g(\mathbf{h})\big(\mathbb{Q}_1(d\mathbf{h}) - \mathbb{Q}_2(d\mathbf{h})\big)\right| : g \in C_{1,1}(\mathcal{H})\right\},$$

where $C_{1,1}(\mathcal{H})$ denotes the set of Lipschitz continuous functions with Lipschitz constant at most 1.

(b) Total variation metric ρ_{tv}:

$$\rho_{tv}(\mathbb{Q}_1, \mathbb{Q}_2) = \sup\left\{|\mathbb{Q}_1(\mathcal{B}) - \mathbb{Q}_2(\mathcal{B})| : \mathcal{B} \in \mathcal{F}(\mathcal{H})\right\}.$$

(c) Hellinger metric ρ_h: Let f_i, $i = 1, 2$ denote the densities of measures $\mathbb{Q}_i$, $i = 1, 2$, with respect to a common dominating measure (e.g. $\mathbb{Q} = (\mathbb{Q}_1 + \mathbb{Q}_2)/2$). Then

$$\rho_h(\mathbb{Q}_1, \mathbb{Q}_2) = \Big(\int_{\mathcal{H}} \big(\sqrt{f_1} - \sqrt{f_2}\big)^2 \mathbb{Q}(d\mathbf{h}) \Big)^{\frac{1}{2}}.$$

(d) Relative entropy distance ρ_e: Let f_i, $i = 1, 2$ denote the densities of measures $\mathbb{Q}_i$, $i = 1, 2$, with respect to a common dominating measure (e.g. $\mathbb{Q} = (\mathbb{Q}_1 + \mathbb{Q}_2)/2$). Then

$$\rho_e(\mathbb{Q}_1, \mathbb{Q}_2) = \int_{\mathcal{H}} f_1(\mathbf{h}) \log \Big(\frac{f_1(\mathbf{h})}{f_2(\mathbf{h})} \Big) d\mathbf{h}$$

The relative entropy distance ρ_e is *not* a metric because it is not symmetric and does not satisfy the triangle inequality.

The following lemma relates the Prohorov metric ρ_p to the other distance functions.

Lemma 5 ([24]) *The distances ρ_w, ρ_h, ρ_{tv} and ρ_e are related to the Prohorov metric as follows.*

(a) Prohorov and Wasserstein metrics: $\rho_p^2 \leq \rho_w \leq (\mathbf{diam}(\mathcal{H}) + 1)\rho_p$, where $\mathbf{diam}(\mathcal{H}) = \sup\{\|\mathbf{h}_1 - \mathbf{h}_2\| : \mathbf{h}_i \in \mathcal{H}, i = 1, 2\}$.

(b) Prohorov and Total variation metrics: $\rho_p \leq \rho_{tv}$

(c) Prohorov and Hellinger metrics: $\rho_p \leq \rho_h$

(d) Prohorov metric and the relative entropy distance: $\rho_p \leq \sqrt{\rho_e/2}$

These bounds imply that for any uncertainty set of the form $\mathcal{Q} = \{\mathbb{Q} : \rho(\mathbb{Q}, \mathbb{Q}_0) \leq \delta\}$, where the metric ρ is given by ρ_w, ρ_{tv}, ρ_h or ρ_e, one can choose $\beta(\delta) > 0$ such that $\mathcal{Q} \subseteq \tilde{\mathcal{Q}} = \{\mathbb{Q} : \rho_p(\mathbb{Q}, \mathbb{Q}_0) \leq \beta(\delta)\}$, i.e. $\tilde{\mathcal{Q}}$ is a *conservative* approximation of $\mathcal{Q}$. Next we introduce the concept of *coupling* of random variables and relate it to the probability metrics.

Definition 3 (Coupling of random variables) *A random variable $\tilde{\mathbf{X}}$ is said to be a* copy *or a* representation *of the random variable $\mathbf{X}$ if and only if they have the same distribution, i.e. $\tilde{\mathbf{X}} \overset{D}{=} \mathbf{X}$. A collection of random variables $\{\tilde{\mathbf{X}}^\alpha : \alpha \in \mathcal{A}\}$ defined on a common probability space $(\Omega, \mathcal{F}(\Omega), \mathbb{P})$ is said to be a* coupling *of the collection $\{\mathbf{X}^\alpha : \alpha \in \mathcal{A}\}$ if and only if $\tilde{\mathbf{X}}^\alpha \overset{D}{=} \mathbf{X}$, for all $\alpha \in \mathcal{A}$.*

Note that only the individual $\tilde{\mathbf{X}}^\alpha$ are copies of the individual $\mathbf{X}^\alpha$, the whole collection is $\{\tilde{\mathbf{X}}^\alpha : \alpha \in \mathcal{A}\}$ is *not* a copy of $\{\mathbf{X}^\alpha : \alpha \in \mathcal{A}\}$, i.e. the joint distribution of $\{\tilde{\mathbf{X}}^\alpha : \alpha \in \mathcal{A}\}$ need not be the same as that of $\{\mathbf{X}^\alpha : \alpha \in \mathcal{A}\}$.

Theorem 4 (Strassen-Dudley) *Let $\mathbf{X}^1 \sim \mathbb{Q}_1$ and $\mathbf{X}^2 \sim \mathbb{Q}_2$ be two random variables taking values in $\mathcal{H}$. Suppose $\rho_p(\mathbb{Q}_1, \mathbb{Q}_2) \leq \beta$. Then there exists a coupling $(\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ of $(\mathbf{X}_1, \mathbf{X}_2)$ such that*

$$\mathbb{P}\Big(\|\tilde{\mathbf{X}}_1 - \tilde{\mathbf{X}}_2\| > \beta\Big) \leq \beta. \tag{17}$$

Proof: This result was established by Strassen [41] for complete separable metric spaces and extended to arbitrary separable metric spaces by Dudley [15]. See also Rachev [34]. ■

This result establishes that if two probability measures $\mathbb{Q}_i$, $i = 1, 2$, are "close" in the Prohorov metric then there exists a coupling $(\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ such that the samples are "close" with high probability. This result can be improved if the random variables $\mathbf{X}_i$, $i = 1, 2$, are bounded w.p.1.

Theorem 5 *Let $\mathbf{X}^1 \sim \mathbb{Q}_1$ and $\mathbf{X}^2 \sim \mathbb{Q}_2$ are two random variables taking values in $\mathcal{H}$. Suppose $\rho_p(\mathbb{Q}_1, \mathbb{Q}_2) \leq \beta$ and $\|\mathbf{X}_i\| \leq R$ a.s., $i = 1, 2$. Then there exists a coupling $(\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2)$ of $(\mathbf{X}_1, \mathbf{X}_2)$ such that*

$$\mathbb{E}\left(\|\tilde{\mathbf{X}}_1 - \tilde{\mathbf{X}}_2\|\right) \leq (1 + 2R)\beta, \tag{18}$$

where the expectation is with respect to the common probability measure $\mathbb{P}$.

Proof: The Wasserstein metric $\rho_w(\mathbb{Q}^1, \mathbb{Q}^2)$ between probability measures $\mathbb{Q}^1$ and $\mathbb{Q}^2$ can be equivalently characterized as follows.

$$\rho_w(\mathbb{Q}_1, \mathbb{Q}_2) = \inf\left\{\mathbb{E}\left[\|\tilde{\mathbf{X}^1} - \tilde{\mathbf{X}^2}\|\right] : \mathbf{X}^i \sim \mathbb{Q}_i, i = 1, 2,\ (\tilde{\mathbf{X}}^1, \tilde{\mathbf{X}}^2) \text{ is a coupling of } (\mathbf{X}^1, \mathbf{X}^2)\right\}.$$

Since $\|\mathbf{X}_i\| \leq R$ a.s., one can without loss of generality assume that $\mathbf{diam}(\mathcal{H}) \leq 2R$. Thus, the bound $\rho_w \leq (\mathbf{diam}(\mathcal{H}) + 1)\rho_p$ together with the characterization above, yields the result. ■

5 Robust chance constrained sets

In this section we show how to construct sampling based approximation for the robust chance constrained set

$$\bar{\mathcal{X}}_\epsilon = \left\{\mathbf{x} \in \mathcal{X} : \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{Q}(\mathbf{H} : f(\mathbf{x}, \mathbf{H}) > 0) \leq \epsilon\right\},$$

where $\mathcal{Q} = \{\mathbb{Q} : \rho_p(\mathbb{Q}, \mathbb{Q}_0) \leq \beta\}$, and ρ_p denotes the Prohorov metric. Note that the bounds in Lemma 5 imply that one can conservatively approximate an uncertainty set defined in terms of any of the metrics discussed in Section 4 by a set defined in terms of the Prohorov metric. The main results of this section are the robust analogs of Lemma 1, Lemma 2 and Lemma 4.

In this section we define $\mathrm{err}(\mathbf{x})$ as follows.

$$\mathrm{err}(\mathbf{x}) = \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{Q}(\mathbf{H} : f(\mathbf{x}, \mathbf{H}) > 0) \tag{19}$$

Thus, $\bar{\mathcal{X}}_\epsilon = \{\mathbf{x} \in \mathcal{X} : \mathrm{err}(\mathbf{x}) \leq \epsilon\}$. Let $\mathbf{H}^0_{1,N} = \{\mathbf{H}^0_1, \ldots, \mathbf{H}^0_N\}$ denote N IID samples drawn according to the central probability measure $\mathbb{Q}_0$. Let $\mathcal{Y}[N, \beta]$ denote the set

$$\mathcal{Y}[N, \beta] = \left\{\mathbf{x} : f(\mathbf{x}, \mathbf{z}) \leq 0, \forall \mathbf{z} \text{ s.t. } \|\mathbf{z} - \mathbf{H}^0_i\| \leq \beta, i = 1, \ldots, N\right\}. \tag{20}$$

As in Section 3, the set $\mathcal{Y}[N, \beta]$ is defined using a *particular* sequence of IID samples $\mathbf{H}^0_{1,N}$ of length N drawn according to the measure $\mathbb{Q}_0$.

Lemma 6 *Fix $\epsilon > 0$. Suppose $\bar{\mathbf{x}} \in \mathcal{X}$ with $\operatorname{err}(\bar{\mathbf{x}}) > \epsilon$. Then, for all $N \geq 1$,*

$$\mathbb{Q}_0^N \left(\mathbf{H}_{1,N}^0 : \bar{\mathbf{x}} \in \mathcal{Y}[N, \beta]\right) \leq e^{-(\epsilon-\beta)N}. \tag{21}$$

Remark 1 *The probability $\mathbb{Q}_0^N(\bar{\mathbf{x}} \in \mathcal{Y}[N, \beta])$ decays exponentially with N only if $\epsilon > \beta$. Thus, uncertainty in the measure manifests itself as a lower bound on the acheivable error probability.*

Proof: Fix $0 < \eta \leq \epsilon$. Since $\operatorname{err}(\bar{\mathbf{x}}) > \epsilon$ we can select $\mathbb{Q}_1 \in \mathcal{Q}$ such that $\mathbb{Q}_1(\mathbf{H} : f(\bar{\mathbf{x}}, \mathbf{H}) > 0) > \epsilon - \eta$. Let $\mathbf{H}^0 \sim \mathbb{Q}_i$, $i = 1, 2$. Since $\rho_p(\mathbb{Q}_1, \mathbb{Q}_0) \leq \beta$, the Strassen-Dudley Representation Theorem implies that there exists a coupling $(\tilde{\mathbf{H}}^1, \tilde{\mathbf{H}}^0)$ of the pair $(\mathbf{H}^1, \mathbf{H}^0)$ such that (17) holds, i.e. $\mathbb{P}\big(\|\tilde{\mathbf{H}}^1 - \tilde{\mathbf{H}}^0\| > \beta\big) \leq \beta$. Let $\big\{(\tilde{\mathbf{H}}_1^1, \tilde{\mathbf{H}}_1^0), \ldots, (\tilde{\mathbf{H}}_N^1, \tilde{\mathbf{H}}_N^0)\big\}$ denote N IID samples of the jointly distributed pair of random vectors $(\tilde{\mathbf{H}}^1, \tilde{\mathbf{H}}^0)$. Let $\tilde{\mathcal{Y}}[N, \beta]$ denote the set

$$\tilde{\mathcal{Y}}[N, \beta] = \left\{\mathbf{x} : f(\mathbf{x}, \mathbf{z}) \leq 0, \forall \mathbf{z} \text{ s.t. } \|\mathbf{z} - \tilde{\mathbf{H}}_i^0\| \leq \beta, i = 1, \ldots, N\right\}. \tag{22}$$

Since $\mathbf{H}^0 \overset{\mathrm{D}}{=} \tilde{\mathbf{H}}^0$ and $\bar{\mathbf{x}}$ is fixed, we have that

$$\mathbb{Q}_0^N \left(\mathbf{H}_{1,N}^0 : \bar{\mathbf{x}} \in \mathcal{Y}[N, \beta]\right) = \mathbb{P}^N \left(\tilde{\mathbf{H}}_{1,N}^0 : \bar{\mathbf{x}} \in \tilde{\mathcal{Y}}[N, \beta]\right). \tag{23}$$

Moreover,

$$\begin{aligned}
\mathbb{P}^N \left(\tilde{\mathbf{H}}_{1,N}^0 : \bar{\mathbf{x}} \in \tilde{\mathcal{Y}}[N, \beta]\right) &= \prod_{i=1}^N \mathbb{P}\left(\tilde{\mathbf{H}}_i^0 : f(\bar{\mathbf{x}}, \mathbf{z}) \leq 0, \forall \|\mathbf{z} - \tilde{\mathbf{H}}_i^0\| \leq \beta\right), \\
&= \left(\mathbb{P}\left(\tilde{\mathbf{H}} : f(\bar{\mathbf{x}}, \mathbf{z}) \leq 0, \forall \|\mathbf{z} - \tilde{\mathbf{H}}^0\| \leq \beta\right)\right)^N. \qquad (24)
\end{aligned}$$

Each term in (24) can be bounded as follows.

$$\begin{aligned}
&\mathbb{P}\left(f(\bar{\mathbf{x}}, \mathbf{z}) \leq 0, \forall \|\mathbf{z} - \tilde{\mathbf{H}}^0\| \leq \beta\right) \\
&\quad = \mathbb{P}\left(f(\bar{\mathbf{x}}, \mathbf{z}) \leq 0, \forall \|\mathbf{z} - \tilde{\mathbf{H}}^0\| \leq \beta, \|\tilde{\mathbf{H}}^1 - \tilde{\mathbf{H}}^0\| \leq \beta\right) \\
&\qquad + \mathbb{P}\left(f(\bar{\mathbf{x}}, \mathbf{z}) \leq 0, \forall \|\mathbf{z} - \tilde{\mathbf{H}}^0\| \leq \beta, \|\tilde{\mathbf{H}}^1 - \tilde{\mathbf{H}}^0\| > \beta\right), \\
&\quad \leq \mathbb{P}\left(f(\bar{\mathbf{x}}, \tilde{\mathbf{H}}^1) \leq 0\right) + \mathbb{P}\left(\|\tilde{\mathbf{H}}^1 - \tilde{\mathbf{H}}^0\| > \beta\right), \qquad (25) \\
&\quad \leq (1 - \epsilon + \eta) + \beta, \qquad (26)
\end{aligned}$$

where (25) follows from the fact that the probability only increases if one removes restrictions, and (26) follows from the bound (17) and the fact that $\tilde{\mathbf{H}}^1 \overset{\mathrm{D}}{=} \mathbf{H}^1$. From (23), (26) and (24), we have

$$\mathbb{Q}_0^N \left(\mathbf{H}_{1,N}^0 : \bar{\mathbf{x}} \in \mathcal{Y}[N, \beta]\right) = \mathbb{P}^N \left(\tilde{\mathbf{H}}_{1,N}^0 : \bar{\mathbf{x}} \in \tilde{\mathcal{Y}}[N, \beta]\right) \leq (1 - \epsilon + \beta + \eta)^N \leq e^{-N(\epsilon-\beta-\eta)}. \tag{27}$$

Since $\eta \leq \epsilon$ was arbitrary, the result follows. ■

Note that we only generate samples according to the central measure $\mathbb{Q}_0$. The coupling is a construct needed to translate the bound on extremal measure that achieves the supremum in the definition of $\operatorname{err}(\mathbf{x})$ to the measure $\mathbb{Q}_0$.

Suppose $\mathbb{Q}(\mathbf{H} : \|\mathbf{H}\| > R) = 0$ for all $\mathbb{Q} \in \mathcal{Q}$. Define the set $\mathcal{Y}_\lambda[N]$ as follows.

$$\mathcal{Y}_\lambda[N] = \big\{\mathbf{x} : f(\mathbf{x}, \mathbf{z}) \leq 0, \forall \mathbf{z} \text{ s.t. } \|\mathbf{z} - \mathbf{H}_i^0\| \leq \lambda\beta, i = 1, \ldots, N\big\} \tag{28}$$

Then Theorem 5 and Markov's inequality implies the following corollary.

Corollary 3 *Fix $\epsilon > 0$ and $\bar{\mathbf{x}} \in \mathcal{X}$ with $\mathrm{err}(\bar{\mathbf{x}}) > \epsilon$. Suppose $\mathbb{Q}(\mathbf{H} : \|\mathbf{H}\| > R) = 0$ for all $\mathbb{Q} \in \mathcal{Q}$. Then*

$$\mathbb{P}\left(\bar{\mathbf{x}} \in \mathcal{Y}_\lambda[N]\right) \leq \Big(1 - \epsilon + \frac{1 + 2R}{\lambda}\Big)^N. \tag{29}$$

Unlike in Lemma 6, here we have a parameter λ that can be controlled to achieve any desired probability of error ϵ.

Next, we establish a robust analog of Lemma 2. We show that if the VC dimension d_f of the function $f(\cdot, \cdot)$ is finite, $\rho_p(\mathbb{Q}, \mathbb{Q}_0) \leq \beta$, and the number of samples $N = \mathcal{O}(\frac{d_f}{\epsilon - \beta})$ (a precise bound is given in Lemma 7),

$$\mathbb{Q}_0^N\left(\mathbf{H}_{1,N}^0 : \mathcal{Y}[N, \beta] \subseteq \mathcal{X}_\epsilon(\mathbb{Q})\right) \geq 1 - \delta.$$

This result should be interpreted as follows. The distribution of the parameters $\mathbf{H}$ is uncertain and is only known to lie in the uncertainty set $\mathcal{Q} = \{\mathbb{Q} : \rho_p(\mathbb{Q}, \mathbb{Q}_0) \leq \beta\}$ and we want to characterize the set of decisions $\mathbf{x}$ that satisfy $\mathbb{Q}(\mathbf{H} : f(\mathbf{x}, \mathbf{H}) > 0) \leq \epsilon$ no matter which probability measure $\mathbb{Q}$ is selected from the uncertainty set $\mathcal{Q}$. The bound above shows that for $N = \mathcal{O}(\frac{d_f}{\epsilon - \beta})$ the set $\mathcal{Y}[N, \beta]$ is a good approximation for $\mathcal{X}_\epsilon(\mathbb{Q})$ for any fixed $\mathbb{Q}$ with high probability.

Lemma 7 *Fix $\delta > 0$, $\epsilon > \beta$ and $\mathbb{Q}_1 \in \mathcal{Q}$. Suppose the VC dimension d_f of the function $f(\cdot, \cdot)$ is finite and $\beta + 2^{-\beta/2} < 1$. Then $\mathbb{Q}_0^N\left(\mathbf{H}_{1,N}^0 : \mathcal{Y}[N, \beta] \not\subseteq \mathcal{X}_\epsilon(\mathbb{Q}_1)\right) \leq \delta$, for all N satisfying*

$$\begin{aligned} N \quad \geq \quad \max\Big\{ \quad & d, \quad \tfrac{8}{\epsilon}, \quad \tfrac{2d_f}{e(1-\beta)} \ln\left(\tfrac{e}{1-\beta}\right) + \tfrac{2}{1-\beta} \ln\left(\tfrac{e}{(e-1)\delta}\right) + 1, \\ & \tfrac{4d_f}{\epsilon-\mu} \log\left(\tfrac{12}{\epsilon-\mu}\right) + \tfrac{4}{\epsilon-\mu} \log\left(\tfrac{2}{\delta(1-\beta)}\right) \Big\}, \end{aligned}$$

where $\mu = 2\big(\frac{\epsilon}{2} + \log(\beta + 2^{-\epsilon/2})\big)$.

Remark 2 *Since $\beta = 0$ implies $\mu = 0$, we recover the non-robust result in Lemma 2 when $\beta = 0$.*

Proof: Since the measure $\mathbb{Q}_1 \in \mathcal{Q}$ is fixed, we will abbreviate $\mathcal{X}_\epsilon(\mathbb{Q}_1)$ by $\mathcal{X}_\epsilon$. Let $\mathcal{X}_\epsilon^c$ denote the complement of the set $\mathcal{X}_\epsilon$. As in the proof of Lemma 6, let $\mathbf{H}^i \sim \mathbb{Q}_i$, $i = 1, 2$ and let $(\tilde{\mathbf{H}}^1, \tilde{\mathbf{H}}^0)$ denote a coupling of the pair $(\mathbf{H}^1, \mathbf{H}^0)$ such that (17) holds, i.e. $\mathbb{P}\big(\|\tilde{\mathbf{H}}^1 - \tilde{\mathbf{H}}^0\| > \beta\big) \leq \beta$. Let $\big\{(\tilde{\mathbf{H}}_1^1, \tilde{\mathbf{H}}_1^0), \ldots, (\tilde{\mathbf{H}}_N^1, \tilde{\mathbf{H}}_N^0)\big\}$ denote N IID samples of the jointly distributed pair of random vectors $(\tilde{\mathbf{H}}^1, \tilde{\mathbf{H}}^0)$. Then

$$\begin{aligned} \mathbb{Q}_0^N\left(\mathbf{H}_{1,N}^0 : \mathcal{Y}[N, \beta] \not\subseteq \mathcal{X}_\epsilon\right) &= \mathbb{P}^N\left(\mathbf{H}_{1,N}^0 : \tilde{\mathcal{Y}}[N, \beta] \cap \mathcal{X}_\epsilon^c \neq \emptyset\right), \\ &= \sum_{j=0}^N \mathbb{P}^N\left(\tilde{\mathcal{Y}}[N, \beta] \cap \mathcal{X}_\epsilon^c \neq \emptyset, |\mathcal{I}| = j\right), \end{aligned}$$

where $\mathcal{I} = \big\{i \in \{1, ..., N\} : \|\tilde{\mathbf{H}}_i^1 - \tilde{\mathbf{H}}_i^0\| \leq \beta\big\}$. For a set $\mathcal{I} \subset \{1, \ldots, N\}$ let $\mathcal{A}(\mathcal{I})$ denote the event

$$\mathcal{A}(\mathcal{I}) = \Big\{(\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1)_{i=1,\ldots,N} : \|\tilde{\mathbf{H}}_k^1 - \tilde{\mathbf{H}}_k^0\| \leq \beta, \forall k \in \mathcal{I}, \|\tilde{\mathbf{H}}_k^1 - \tilde{\mathbf{H}}_k^0\| \leq \beta, \forall k \notin \mathcal{I}\Big\}$$

and let $\mathcal{Y}[\mathcal{I}, \beta] = \big\{\mathbf{x} : f(\mathbf{x}, \mathbf{z}) \leq 0, \forall \mathbf{z} \text{ s.t. } \|\mathbf{z} - \tilde{\mathbf{H}}_i^0\| \leq \beta, i \in \mathcal{I}_j\big\}$. Fix $\mathcal{I}_1, \mathcal{I}_2 \subseteq \{1, ..., N\}$ with $|\mathcal{I}_1| = |\mathcal{I}_2|$. Since $\{(\tilde{\mathbf{H}}_i{}^1, \tilde{\mathbf{H}}_i{}^0)\}$, $i = 1, ..., N$ are IID, it is clear that

$$\begin{aligned} &\mathbb{P}^N\Big((\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1)_{i=1,\ldots,N} : \tilde{\mathcal{Y}}[N, \beta] \cap \mathcal{X}_\epsilon^c \neq \emptyset, \mathcal{A}(\mathcal{I}_1)\Big) \\ &\quad = \mathbb{P}^N\Big((\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1)_{i=1,\ldots,N} : \tilde{\mathcal{Y}}[N, \beta] \cap \mathcal{X}_\epsilon^c \neq \emptyset, \mathcal{A}(\mathcal{I}_2)\Big). \end{aligned} \tag{30}$$

Set $\mathcal{I}_0 = \emptyset$, and $\mathcal{I}_j = \{1, \ldots, j\}$, $j = 1, \ldots, N$. Since there are $\binom{N}{j}$ possible selections for the set $\mathcal{I}_j$ of cardinality j, (30) implies that

$$\begin{aligned} &\sum_{j=0}^N \mathbb{P}^N\Big((\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1)_{i=1,\ldots,N} : \tilde{\mathcal{Y}}[N, \beta] \cap \mathcal{X}_\epsilon^c \neq \emptyset, |\mathcal{I}| = j\Big) \\ &= \sum_{j=0}^N \binom{N}{j} \mathbb{P}^N\Big((\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1)_{i=1,\ldots,N} : \tilde{\mathcal{Y}}[N, \beta] \cap \mathcal{X}_\epsilon^c, \mathcal{A}(\mathcal{I}_j)\Big), \\ &\leq \sum_{j=0}^N \binom{N}{j} \mathbb{P}^N\Big((\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1)_{i=1,\ldots,N} : \mathcal{X}_\epsilon^c \cap \mathcal{Y}[\mathcal{I}_j, \beta] \neq \emptyset, \mathcal{A}(\mathcal{I}_j)\Big) && (31) \\ &= \sum_{j=0}^N \binom{N}{j} \mathbb{P}^j\Big((\tilde{\mathbf{H}}_k^0, \tilde{\mathbf{H}}_k^1)_{k\in\mathcal{I}_j} : \mathcal{X}_\epsilon^c \cap \mathcal{Y}[\mathcal{I}_j, \beta] \neq \emptyset, \mathcal{A}(\mathcal{I}_j)\Big) \cdot \\ &\qquad\qquad \mathbb{P}^{N-j}\Big((\tilde{\mathbf{H}}_k^0, \tilde{\mathbf{H}}_k^1)_{k\notin\mathcal{I}_j} : \|\tilde{\mathbf{H}}_k{}^{\bar{\alpha}} - \tilde{\mathbf{H}}_k{}^0\| > \beta, \forall k \notin \mathcal{I}_j\Big), && (32) \\ &\leq \sum_{j=0}^N \binom{N}{j} \beta^{N-j} \mathbb{P}^j\Big((\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1)_{i=1,\ldots,j} : \mathcal{X}_\epsilon^c \cap \mathcal{Y}[\mathcal{I}_j, \beta] \neq \emptyset, \mathcal{A}(\mathcal{I}_j)\Big) && (33) \\ &\leq \sum_{j=0}^N \binom{N}{j} \beta^{N-j} \mathbb{P}^j\Big(\tilde{\mathbf{H}}_{1,j}^1 : \exists \mathbf{x} \in \mathcal{X}_\epsilon^c \text{ s.t. } f(\mathbf{x}, \tilde{\mathbf{H}}_k^1) \leq 0, \forall k \in \mathcal{I}_j\Big), && (34) \end{aligned}$$

where (31) and (34) follows from the fact that the probability only increases if one removes restrictions, (32) follows from the fact that $\{(\tilde{\mathbf{H}}_i{}^1, \tilde{\mathbf{H}}_i{}^0)\}$, $i = 1, ..., N$ are IID, and (33) follows from the bound (17). Note that the bound (34) only involves the random vector $\mathbf{H}^1$, or equivalently the (unknown) true measure $\mathbb{Q}_1$. Thus, once again we have used coupling to translate a bound in terms of the central measure $\mathbb{Q}_0$ to one involving the measure $\mathbb{Q}_1$. We do not need coupling beyond this stage of the proof. In the rest of this proof we bound (34) using Lemma 2 applied to the (unknown) measure $\mathbb{Q}_1$. Let $N_0 = \lfloor \frac{8}{\epsilon} \rfloor$. Then

$$\begin{aligned}
&\sum_{j=0}^{N} \binom{N}{j} \beta^{N-j} \mathbb{P}^j \left(\tilde{\mathbf{H}}^1_{1,j} : \exists \mathbf{x} \in \mathcal{X}^c_\epsilon \text{ s.t. } f(\mathbf{x}, \tilde{\mathbf{H}}^1_k) \leq 0, \forall k \in \mathcal{I}_j \right) \\
&= \sum_{j=0}^{N_0} \binom{N}{j} \beta^{N-j} \mathbb{P}^j \left(\tilde{\mathbf{H}}^1_{1,j} : \exists \mathbf{x} \in \mathcal{X}^c_\epsilon \text{ s.t. } f(\mathbf{x}, \tilde{\mathbf{H}}^1_k) \leq 0, \forall k \in \mathcal{I}_j \right) \\
&\quad + \sum_{j=N_0+1}^{N} \binom{N}{j} \beta^{N-j} \mathbb{P}^j \left(\tilde{\mathbf{H}}^1_{1,j} : \exists \mathbf{x} \in \mathcal{X}^c_\epsilon \text{ s.t. } f(\mathbf{x}, \tilde{\mathbf{H}}^1_k) \leq 0, \forall k \in \mathcal{I}_j \right), \\
&\leq \sum_{j=0}^{N_0} \binom{N}{j} \beta^{N-j} + \sum_{j=N_0+1}^{N} \binom{N}{j} \beta^{N-j} \mathbb{P}^j \left(\tilde{\mathbf{H}}^1_{1,j} : \exists \mathbf{x} \in \mathcal{X}^c_\epsilon \text{ s.t. } f(\mathbf{x}, \tilde{\mathbf{H}}^1_k) \leq 0, \forall k \in \mathcal{I}_j \right) \\
&\leq \sum_{j=0}^{N_0} \binom{N}{j} \beta^{N-j} + \sum_{j=N_0+1}^{N} \binom{N}{j} \beta^{N-j} \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2}, \qquad (35)
\end{aligned}$$

where (35) follows from Lemma 2 and the bound (13). The rest of this proof is tedious algebra to prove a "nice" bound on (35).

$$\begin{aligned}
&\sum_{j=0}^{N_0} \binom{N}{j} \beta^{N-j} + \sum_{j=N_0+1}^{N} \binom{N}{j} \beta^{N-j} \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2} \\
&= \underbrace{\sum_{j=0}^{N_0} \binom{N}{j} \beta^{N-j} \left(1 - \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2} \right)}_{\tau_1} + \underbrace{\sum_{j=0}^{N} \binom{N}{j} \beta^{N-j} \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2}}_{\tau_2} \qquad (36)
\end{aligned}$$

To complete the proof we show that if N is large enough the terms τ_1 and τ_2 are bounded by $\tau_1 \leq \delta\beta$ and $\tau_2 \leq \delta(1-\beta)$, which implies that $\tau_1 + \tau_2 \leq \delta$. We can bound τ_1 as follows. Let $d_0 = \lfloor \frac{d_f}{e} \rfloor$ where e is the base of natural logarithm. Then

$$\tau_1 = \sum_{j=0}^{d_0} \binom{N}{j} \beta^{N-j} \left(1 - \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2} \right) + \sum_{j=d_0+1}^{N_0} \binom{N}{j} \beta^{N-j} \left(1 - \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2} \right).$$

Note that for $\frac{d_f}{e} \leq d_0 + 1 \leq j \leq N_0 \leq \frac{8}{\epsilon}$. Thus, we have

$$\begin{aligned}
1 - \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2} &\leq 1 - \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon N_0/2}, \\
&\leq 1 - \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon \frac{8}{2\epsilon}}, \\
&= 1 - \left(\frac{2ej}{d_f} \right)^{d_f} 2^{-3}, \\
&\leq 1 - \left(\frac{2ed_f}{d_f e} \right)^{d_f} 2^{-3}, \\
&= 1 - 2^{d_f-3} \leq 0.
\end{aligned}$$

The last inequality follows from the assumption that $d_f > 3$. Therefore,

$$\begin{aligned}
\tau_1 &\leq \sum_{j=0}^{d_0} \binom{N}{j} \beta^{N-j} \left(1 - \left(\frac{2ej}{d} \right)^d 2^{1-\epsilon j/2} \right), \\
&\leq \sum_{j=0}^{d_0} \binom{N}{j} \beta^{N-j}, \\
&\leq \frac{\beta N}{N-d_0} \sum_{j=0}^{d_0} \binom{N-1}{j} \beta^{N-1-j}, \\
&\leq \frac{N\beta(1-\beta)^{-d_0}}{N-d_0} \sum_{j=0}^{d_0} \binom{N-1}{j} \beta^{N-1-j}(1-\beta)^j, \\
&= \left(\frac{N\beta(1-\beta)^{-d_0}}{N-d_0} \right) P(1-\beta, N-1, d_0), \qquad (37)
\end{aligned}$$

where $P(p, N, s)$ denotes the probability of at most s successes in N IID Bernoulli trials, each with a success probability p. Let $\theta = 1 - \frac{d_0}{(N-1)(1-\beta)}$. Then, Chernoff bound implies that

$$\tau_1 \leq \frac{N\beta(1-\beta)^{-d_0}}{N-d_0} \exp\left\{ -\frac{(N-1)(1-\beta)}{2} + d_0 \right\}.$$

For $N \geq d_f \geq ed_0$ we have $\frac{N}{N-d_0} \leq \frac{e}{e-1}$. Therefore,

$$\tau_1 \leq \frac{e\beta(1-\beta)^{-d_f/\epsilon}}{e-1} \exp\left\{ -\frac{(N-1)(1-\beta)}{2} + \frac{d}{\epsilon} \right\}. \qquad (38)$$

Thus, $\tau_1 \leq \delta\beta$ for all

$$N \geq \frac{2d_f}{e(1-\beta)} \ln\left(\frac{e}{1-\beta} \right) + \frac{2}{1-\beta} \ln\left(\frac{e}{(e-1)\delta} \right) + 1 \qquad (39)$$

Next, we bound τ_2 as follows.

$$\begin{aligned}
\tau_2 &= \sum_{j=0}^{N} \binom{N}{j} \beta^{N-j} \left(\frac{2ej}{d_f} \right)^{d_f} 2^{1-\epsilon j/2}, \\
&= 2\left(\frac{2e}{d_f} \right)^{d_f} \sum_{j=0}^{N} \binom{N}{j} j^{d_f} \beta^{N-j} 2^{-\epsilon j/2}, \\
&\leq 2\left(\frac{2e}{d_f} \right)^{d_f} N^{d_f} \sum_{j=0}^{N} \binom{N}{j} \beta^{N-j} 2^{-\epsilon j/2}, \\
&\leq 2\left(\frac{2e}{d_f} \right)^{d_f} N^{d_f} (\beta + 2^{-\epsilon/2})^N, \qquad (40)
\end{aligned}$$

Since $\beta+2^{-\epsilon/2} \leq \beta+2^{-\beta/2} < 1$, $\mu = 2(\frac{\epsilon}{2}+\log(\beta+2^{-\epsilon/2}))$ is well defined. Then an analysis similar to the one given in [9] (see also [1] pg. 95 for details) shows that

$$\tau_2 \leq 2\left(\frac{2e}{d_f}\right)^{d_f} N^{d_f}(2^{-(\epsilon-\mu)/2})^N$$

Thus, $\tau_2 \leq (1-\delta)\beta$ for all

$$N \geq \frac{4d_f}{\epsilon-\mu}\log\left(\frac{12}{\epsilon-\mu}\right)+\frac{4}{\epsilon-\mu}\log\left(\frac{2}{\delta(1-\beta)}\right) \tag{41}$$

The result follows from (36), (39), and (41). ■

The last result in this section is the robust analog of Lemma 4.

Lemma 8 *Fix $\epsilon > 0$. Let $\widehat{\mathbf{x}}$ denote the optimal solution of the robust sampled problem (8). Then $\mathbb{Q}_0^N(\mathbf{H}_{1,N}^0 : \widehat{\mathbf{x}} \notin \bar{\mathcal{X}}_\epsilon) \leq \left(\frac{eN}{n}\right)^n e^{-(\epsilon-\beta)(N-n)}$.*

Proof: The robust chance constrained problem (8) has constraints of the form

$$f(\mathbf{x},\mathbf{z}) \leq 0, \quad \|\mathbf{z}-\mathbf{H}_i^0\| \leq \beta, \quad i=1,\ldots,N.$$

Suppose a constraint of the form $f(\mathbf{x},\bar{\mathbf{z}}) \leq 0$ is a support constraint for the robust chance constrained problem (8). We will associate this support constraint with $k = \text{argmin}\left\{i : \|\bar{\mathbf{z}}-\mathbf{H}_i^0\| \leq \beta\right\}$.

Let $\mathcal{I} \subseteq \{1,\ldots,N\}$ with $|I| \leq n$ and let

$$\mathcal{H}_{\mathcal{I}}^N = \{(\mathbf{h}_1,\ldots,\mathbf{h}_N) : \text{all support constraints are associated with some } i \in \mathcal{I}\}.$$

Then Theorem 3 implies $\mathcal{H}^N = \cup_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|\leq n\}}\mathcal{H}_{\mathcal{I}}^N$. Thus,

$$\begin{aligned}
&\mathbb{Q}_0^N(\mathbf{H}_{1,N} : \widehat{\mathbf{x}} \notin \bar{\mathcal{X}}_\epsilon)\\
&\quad= \sum_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|\leq n\}} \mathbb{Q}_0^N(\mathbf{H}_{1,N} \in \mathcal{H}_{\mathcal{I}}^N : \widehat{\mathbf{x}}_{\mathcal{I}} \notin \bar{\mathcal{X}}_\epsilon)\\
&\quad= \sum_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|= n\}} \Big(\mathbb{Q}_0^n(\mathbf{H}_{i\in I} : \widehat{\mathbf{x}}_{\mathcal{I}} \notin \bar{\mathcal{X}}_\epsilon)\prod_{i\notin\mathcal{I}}\mathbb{Q}_0(\mathbf{H}_i : f(\widehat{\mathbf{x}}_{\mathcal{I}},\mathbf{z}) \leq 0,\ \forall\|\mathbf{z}-\mathbf{H}_i\| \leq \beta|\mathcal{A}_{\mathcal{I}})\Big),
\end{aligned}$$

where $\widehat{\mathbf{x}}_{\mathcal{I}}$ denotes the optimal solution of the robust sampled problem (5) with only the robust constraints corresponding to the samples $i \in \mathcal{I}$ present, $\mathcal{A}_{\mathcal{I}}$ is the event $\mathcal{A}_{\mathcal{I}} = \{\mathbf{H}_{i\in I} : \widehat{\mathbf{x}}_{\mathcal{I}} \notin \bar{\mathcal{X}}_\epsilon\}$ and each term in the sum can be written as the product because $\mathbf{H}_{1,N}^0$ are IID samples. Lemma 1 implies that $\mathbb{Q}_0(\mathbf{H} : f(\widehat{\mathbf{x}}_{\mathcal{I}},\mathbf{z}) \leq 0,\ \forall\|\mathbf{z}-\mathbf{H}\| \leq \beta|\mathcal{A}_{\mathcal{I}}) \leq e^{-(\epsilon-\beta)}$. Thus,

$$\begin{aligned}
\mathbb{Q}_0^N(\mathbf{H}_{1,N} : \widehat{\mathbf{x}} \notin \mathcal{X}_\epsilon) &\leq e^{-(\epsilon-\beta)(N-n)} \sum_{\{\mathcal{I}\subseteq\{1,\ldots,N\}:|\mathcal{I}|= n\}} \mathbb{Q}_0^n(\mathbf{H}_{i\in I} : \widehat{\mathbf{x}}_{\mathcal{I}} \notin \bar{\mathcal{X}}_\epsilon),\\
&\leq e^{-(\epsilon-\beta)(N-n)}\Big(\sum_{k=1}^n \binom{N}{k}\Big) \leq \Big(\frac{eN}{n}\Big)^n e^{-(\epsilon-\beta)(N-n)},
\end{aligned}$$

where the last inequality follows from the bound (13). ■

6 Tractability of the robust sampled problem

In Section 1 we introduced the robust sampled problem (8) as an approximation for the ambiguous chance constrained problem (6) and in Section 5 we established bounds on the number of samples N required to approximate the robust feasible set $\bar{\mathcal{X}}_\epsilon$ and also on the number of samples required to only ensure that the optimal solution $\widehat{\mathbf{x}}$ of the robust problem (8) is feasible for (6). All along we have implicitly assumed that the robust sampled problem (8) is efficiently solvable. In this section, we characterize the functions $f(\cdot,\cdot)$, the probability metric ρ and the norm $\|\cdot\|$ on the parameter space $\mathcal{H}$ for which the robust sampled problem (8) is tractable both in theory and in practice. The results in this section are motivated by [7].

We restrict attention to the following two classes of constraint functions.

(a) Affine functions: $\mathcal{X} = \mathbf{R}^n$, $\mathcal{H} = \mathbf{R}^{n+1}$, and $f(\mathbf{x},(h_0,\mathbf{h})) = h_0 + \mathbf{h}^T\mathbf{x}$.

(b) Second-order cone functions: $\mathbf{x} \in \mathbf{R}^n$, $\mathcal{H} = \big\{\mathbf{h} = (\mathbf{A},\mathbf{b},\mathbf{u},v) : \mathbf{A} \in \mathbf{R}^{p\times n}, \mathbf{b} \in \mathbf{R}^p, \mathbf{u} \in \mathbf{R}^n, v \in \mathbf{R}\big\}$, and $f(\mathbf{x},\mathbf{h}) = \sqrt{(\mathbf{A}\mathbf{x}+\mathbf{b})^T(\mathbf{A}\mathbf{x}+\mathbf{b})} - \mathbf{u}^T\mathbf{x} - v$.

The uncertainty set $\mathcal{Q}$ considered in this paper is given by $\mathcal{Q} = \{\mathbb{Q} : \rho_p(\mathbb{Q},\mathbb{Q}_0) \le \beta\}$ where ρ_p denotes the Prohorov metric. Since the Prohorov metric is defined in terms of the norm $\|\cdot\|$ on the space $\mathcal{H}$, we first select this norm. We restrict attention to norms that satisfy

$$\|\mathbf{u}\| = \|\,|\mathbf{u}|\,\|, \tag{42}$$

where $|\mathbf{u}|$ denotes the vector obtained by taking the absolute value of each of the components. For a given norm $\|\cdot\|$, the constant β defining the uncertainty set $\mathcal{Q}$ is set by the desired level of confidence. Note that β can also be set in terms of any distance measure that is an upper bound for the Prohorov metric. See Section 4 for details.

First we consider the case of affine constraint functions $f(\mathbf{x},\mathbf{h}) = h_0 + \mathbf{h}^T\mathbf{x}$. Let $\mathbf{e}_j$, $j = 1,\ldots,n+1$ denote the j-th basis vector in $\mathbf{R}^{n+1}$. Define $\mathcal{U}(\mathbf{h}) = \big\{\mathbf{z} : \mathbf{z} = \mathbf{h} + \sum_{i=1}^{n+1} w_j\mathbf{e}_j, \|\mathbf{w}\| \le \beta\big\}$. Then the robust sampled problem (8) is given by

$$\begin{array}{ll} \min & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} & z_0 + \mathbf{z}^T\mathbf{x} \le 0, \quad \forall \mathbf{z} \in \mathcal{U}(\mathbf{H}_i^0), \quad i = 1,\ldots,N, \\ & \mathbf{x} \in \mathcal{X}. \end{array} \tag{43}$$

Results in [7] show that (43) can be reformulated as follows.

$$\begin{array}{llll} \min & \mathbf{c}^T\mathbf{x} & & \\ \text{s.t.} & f(\mathbf{x},\mathbf{H}_i^0) \le -\beta y_i, & i = 1,\ldots,N, & \\ & |x_j| \le t_j^i, & j = 1,\ldots,n, & i = 1,\ldots,N, \\ & 1 \le t_{n+1}^i, & i = 1,\ldots,N, & \\ & \|\mathbf{t}^i\|_* \le y_i, & i = 1,...,N, & \\ & \mathbf{y} \in \mathbf{R}^N,\ \mathbf{t}^i \in \mathbf{R}^{n+1}, & i = 1,\ldots,N, & \\ & \mathbf{x} \in \mathcal{X}. & & \end{array} \tag{44}$$

where $\|\mathbf{s}\|_* = \max_{\{\|\mathbf{r}\|\leq 1\}}\{\mathbf{s}^T\mathbf{r}\}$ denotes the dual norm of $\|\cdot\|$. For the $\mathcal{L}_1$ or $\mathcal{L}_\infty$ norms (44) reduces to a linear program; whereas when the norm $\|\cdot\|$ is an $\mathcal{L}_p$-norm, $p \neq \{1,\infty\}$, (44) is equivalent to a second-order cone program.

Next, consider the same of the second-order cone constraints. Let $\mathbf{e}_j \in \mathbf{R}^{(p+1)(n+1)}$ denote the j-th standard basis vector in $\mathbf{R}^{(p+1)(n+1)}$. For $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n] \in \mathbf{R}^{p\times n}$ let

$$\text{vec}(\mathbf{A}) = \begin{bmatrix}\mathbf{a}_1^T & \cdots & \mathbf{a}_n^T\end{bmatrix}^T \in \mathbf{R}^{pn},$$

and, for $\mathbf{h} \in \mathbf{R}^{(p+1)(n+1)}$, define $\mathcal{U}(\mathbf{h}) = \big\{\mathbf{z} : \mathbf{z} = \mathbf{h} + \sum_{j=1}^{(p+1)(n+1)} w_j\mathbf{e}_j,\ \|\mathbf{w}\| \leq \beta\big\}$. It is shown in [7] that any feasible solution to the problem (45) below is also feasible for the robust sampled problem (8).

$$\begin{array}{lll}
\min & \mathbf{c}^T\mathbf{x} & \\
\text{s.t.} & f(\mathbf{x}, \mathbf{H}_i^0) \leq -\beta y_i, & i = 1,\ldots,N, \\
& g_j^i(\mathbf{x}) \leq t_j^i, & j = 1,\ldots,(p+1)(n+1), \quad i = 1,\ldots,N, \\
& \|\mathbf{t}^i\|_* \leq y_i, & i = 1,\ldots,N, \\
& \mathbf{y} \in \mathbf{R}^N, \mathbf{t}^i \in \mathbf{R}^{(p+1)(n+1)}, & i = 1,\ldots,N, \\
& \mathbf{x} \in \mathcal{X}, &
\end{array} \tag{45}$$

where

$$g_j^i(\mathbf{x}) = \begin{cases} |x_l| & j = p(l-1)+k,\ k = 1,\ldots,p, \quad l = 1,\ldots,n, \\ 1 & j = pn+k, \quad k = 1,\ldots,p \\ |x_l| & j = (p+1)n+l, \quad l = 1,\ldots,n \\ 1 & j = (p+1)(n+1), \end{cases}$$

The problem (45) is a second-order cone program for all $\mathcal{L}_p$ norms.

7 Conclusion

In this paper we extend the sample complexity results known for chance constrained problems to ambiguous chance constrained problems where the uncertainty set is given by a ball defined in terms of the Prohorov metric. We approximate the ambiguous chance constrained problem by a robust sampled problem where each constraint is a robust constraint centered at a sample drawn according to the center of the uncertainty set defining the ambiguous chance constrained problem. The main contribution of this paper is to show that the robust sampled problem is a good approximation for the ambiguous chance constrained problem with high probability. Our extensions are based on the Strassen-Dudley Representation Theorem that states that when the *distributions* of two random variables are close in the Prohorov metric one can construct a *coupling* of the random variables such that the *samples* are close with high probability. Coupling is just a construct needed to prove the results; it is never used in computing the solution to the robust sampled problem.

The results in this paper should be viewed as a first step towards solving ambiguous chance constrained problems. Several issues still remain unresolved. We only consider uncertainty sets that are norm balls defined in terms of the Prohorov metric. One could consider "tiling" a more general uncertainty set by norm balls of a given radius and construct a robust sampled problem by drawing samples according to the centers of the balls (a simplified version of this idea appears in [11]). Since the constant ϵ that controls the violation probability in the ambiguous chance constrained problem has to be greater that the radius β of the norm ball, such an approach is attractive even when the uncertainty is a norm ball. However, it is not clear how to efficiently select the centers of the balls to "tile" the uncertainty set.

In Section 4 we introduce several probability metrics and show that uncertainty set $\mathcal{Q} = \{\mathbb{Q} : \rho(\mathbb{Q}, \mathbb{Q}_0) \leq \delta\}$ can be conservatively approximated by a uncertainty set $\tilde{\mathcal{Q}} = \{\mathbb{Q} : \rho_p(\mathbb{Q}, \mathbb{Q}_0) \leq \delta(\beta)\}$ defined in terms of the Prohorov metric. However, we have no way of measuring the "blow-up" of the uncertainty set that occurs in changing the metrics. This issue can be resolved by either establishing tight lower bounds on the Prohorov metric or by constructing Representation results for the other metrics. Ideally one would like to get logarithmically separated upper and lower bounds as in [32, 33].

Finally, there is the issue of proving worst-case lower bounds on the number of samples required to learn the solution of an ambiguous chance constrained problem, i.e. a refinement of Lemma 3 that accounts for ambiguity in the measure.

References

[1] M. Anthony and N. Biggs. *Computational Learning Theory.* Cambridge University Press, 1992.

[2] P. Artzner, F. Delbaen, J-P. Eber, and D. Heath. Coherent risk measures. *Math. Finance*, 9(203-228), 1999.

[3] A. Ben-Tal and A. Nemirovski. Robust truss topology design via semidefinite programming. *SIAM J. Optim.*, 7(4):991–1016, 1997.

[4] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Math. Oper. Res.*, 23(4):769–805, 1998.

[5] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Oper. Res. Lett.*, 25(1):1–13, 1999.

[6] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization.* SIAM, Philadelphia, PA, 2001.

[7] D. Bertsimas and M. Sim. Robust conic optimization. Under review in *Math. Prog.*, May 2004.

[8] J. R. Birge and J. R. B. Wets. Designing approximation schemes for stochastic optimization problems, in particular for stochastic programs with recourse. *Math. Prog. Study*, 27(54-102), 1986.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, Oct 1989.

[10] G. Calafiore and M. C. Campi. Uncertain convex programs: Randomized solutions and confidence levels. To appear in *Math. Prog.*, 2003.

[11] G. Calafiore and M. C. Campi. Decision making in an uncertain environment: the scenario-based optimization approach. Working paper, 2004.

[12] A. Charnes and W. W. Cooper. Uncertain convex programs: randomized solutions and confidence levels. *Mgmt. Sc.*, 6:73–79, 1959.

[13] Z. Chen and L.G. Epstein. Ambiguity, risk and asset returns in continuous time. Mimeo, 2000.

[14] D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. To appear in *Math. Oper. Res.*, 2001.

[15] R. M. Dudley. Distance of probability measures and random variables. *Ann. Math. Stat.*, 39:1563–1572, 1968.

[16] J. Dupačová. The minimax approach to stochastic program and illustrative application. *Stochastics*, 20:73–88, 1987.

[17] J. Dupačová. Stochastic programming: minimax approach. In *Encyclopedia of Optimization.* Kluwer, 2001.

[18] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–1064, 1997.

[19] L. El Ghaoui and A. Nilim. The curse of uncertainty in dynamic programming and how to fix it. To appear in *Oper. Res.*. UC Berkeley Tech Report UCB-ERL-M02/31, November 2002.

[20] L. El Ghaoui, F. Oustry, and H. Lebret. Robust solutions to uncertain semidefinite programs. *SIAM J. Optim.*, 9(1):33–52, 1998.

[21] L. G. Epstein and M. Schneider. Recursive multiple priors. Technical Report 485, Rochester Center for Economic Research, August 2001. Available at http://rcer.econ.rochester.edu. To appear in *J. Econ. Theory.*

[22] L. G. Epstein and M. Schneider. Learning under Ambiguity. Technical Report 497, Rochester Center for Economic Research, October 2002. Available at http://rcer.econ.rochester.edu.

[23] H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Fin. and Stoch.*, 6:429–447, 2002.

[24] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *Intl. Stat. Rev.*, 7(3):419–435, 2002.

[25] I. Gilboa and D. Schmeidler. Maxmin expected utility with non-unique prior. *J. Math. Econ.*, 18(2):141–153, 1989.

[26] D. Goldfarb and Iyengar G. Robust portfolio selection problems. *Math. Oper. Res.*, 28(1):1–38, 2003.

[27] L. P. Hansen and T. J. Sargent. Robust control and model uncertainty. *American Economic Review*, 91:60–66, 2001.

[28] G. Iyengar. Robust dynamic programming. To appear in *Math. Oper. Res.*. Available at http://www.corc.ieor.columbia.edu/reports/techreports/tr-2002-07.pdf, 2002.

[29] R. Jagannathan. Minimax procedure for a class of linear programs under uncertainty. *Oper. Res.*, 25:173–177, 1977.

[30] M. J. Kearns and U. V. Vazirani. *An introduction to computational learning theory.* MIT Press, Cambridge, MA, 1997.

[31] P. M. Long. The complexity of learning according to two models of a drifting environment. *Machine Learning*, 37(3):337–354, 1999.

[32] A. Nemirovski. On tractable approximations of randomly perturbed convex constraints. In *Proc. 42nd IEEE Conf. Dec. Contr. (CDC)*, volume 3, pages 2419– 2422, 2003.

[33] A. Nemirovski and A. Shapiro. Scenario approximations of chance constraints. To appear in *Probabilistic and randomized methods for design under uncertainty*, 2004.

[34] S. T. Rachev. *Probability metrics and the stability of stochastic models.* John Wiley & Sons, 1991.

[35] A. Ruszczynski and A. Shapiro, editors. *Stochastic Programming.* Handbook in Operations Research and Management Science. Elsevier, 2003.

[36] A. Ruszczynski and A. Shapiro. Optimization of risk measures. Available at http://www.optimization-online.org/DB_HTML/2004/02/822.html, 2004.

[37] A. Ruszczynski and A. Shapiro. Optimization of risk measures. Available at http://ideas.repec.org/p/wpa/wuwpri/0407002.html, 2004.

[38] A. Shapiro. Some recent developments in stochastic programming. *ORB Newsletter*, 13, March 2004. Available at http://www.ballarat.edu.au/ard/itms/CIAO/ORBNewsletter/issue13.shtml#11.

[39] A. Shapiro and S. Ahmed. On a class of minimax stochastic programs. To appear in *SIAM J. Opt.*, 2004.

[40] A. Shapiro and A. J. Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software.*

[41] V. Strassen. The existence of probability measures with given marginals. *Ann. Math. Stat.*, 36:423–439, 1965.

[42] H. Thorisson. *Coupling, Stationary, and Regeneration.* Probability and its Applications. Springer-Verlag, 2000.

[43] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, NY, 1995.

[44] J. Žáčková. On minimax solutions of stochastic linear programs. *Čas. Pěst. Mat.*, pages 423–430, 1966.