On the connection between compression learning and scenario based optimization ☆

Kostas Margellos[a,*], Maria Prandini[b], John Lygeros[c]

[a] *Department of Industrial Engineering and Operations Research, UC Berkeley, Sutardja Dai Hall 330, Berkeley CA 94720, United States*

[b] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano 20133, Italy*

[c] *Department of Information Technology and Electrical Engineering, ETH Zürich, Physikstrasse 3, Zürich 8092, Switzerland*

Abstract

We investigate the connections between compression learning and scenario based optimization. We first show how to strengthen, or relax the consistency assumption at the basis of compression learning and study the learning and generalization properties of the algorithm involved. We then consider different constrained optimization problems affected by uncertainty represented by means of scenarios. We show that the issue of providing guarantees on the probability of constraint violation reduces to a learning problem for an appropriately chosen algorithm that enjoys compression learning properties. The compression learning perspective provides a unifying framework for scenario based optimization and allows us to revisit the scenario approach and the probabilistically robust design, a recently developed technique based on a mixture of randomized and robust optimization, and to extend the guarantees on the probability of constraint violation to cascading optimization problems.

Keywords: Compression learning, consistent algorithms, randomized optimization, scenario approach, statistical learning theory.

1. Introduction

Optimal decision making in the presence of uncertainty is important for the efficient and economic operation of systems affected by endogenous, or exogenous uncertainties. One approach to deal with uncertainty is through robust optimization. In this case a decision is made such that the constraints are satisfied for all admissible values of the uncertainty [2]. Tractability of the developed techniques relies heavily on the geometry of the uncertainty set. On the other hand, chance constrained optimization allows for constraint violation but with an a-priori specified probability [3], [4]. In [5], [6], different approximations to chance constrained optimization problems are proposed

☆ Research was supported by the European Commission under the projects MoVeS and SPEEDD. The authors would like to thank Prof. Simone Garatti for stimulating discussions and for bringing the work in reference [1] to our attention.

* Corresponding author

Email addresses: kostas.margellos@berkeley.edu (Kostas Margellos), prandini@elet.polimi.it (Maria Prandini), lygeros@control.ee.ethz.ch (John Lygeros)

for the case where the constraints exhibit a specific structure with respect to the uncertainty and under certain assumptions on the underlying probability distribution.

In many cases, however, we are only provided with data, e.g. historical values of the uncertainty. Therefore, research has been devoted towards the development of a data driven decision making paradigm. Under such a set-up, an alternative to robust optimization is scenario based optimization, that involves solving an optimization problem whose constraints depend only on a finite number of uncertainty instances called "scenarios". Scenario based optimization does not require any specific assumption on the probability distribution of the uncertainty or the way in which the uncertainty enters the optimization problem. On the other hand, it does require certain structure of the underlying optimization problem to ensure that the properties of the solution generalize to unseen uncertainty instances and hence to provide guarantees regarding the probability of constraint satisfaction. For problems that are convex with respect to the decision variables the so called scenario approach [7], [8], [9], offers an already mature theoretical framework for analyzing the generalization properties of the optimal solution. In the non-convex case, tools from statistical learning [10], [11], [12] based on the VC theory can be employed to provide guarantees on the probability of constraint satisfaction for any feasible solution of an optimization problem [13], [14], [15].

In this paper we explore the links between learning theory and the scenario approach to scenario based optimization without resorting to VC theoretic results. To this end we exploit the results of [1] and consider compression learning algorithms, that are based on an alternative notion of learning under an assumption referred to as consistency. We first show how using ideas from the scenario approach theory one can strengthen or relax the consistency assumption, and analyze the resulting learnability properties. We then return to optimization problems and show that the problem of providing guarantees regarding the probability of constraint violation can be equivalently thought of as a learning problem for an appropriately chosen algorithm that enjoys some compression property. Different classes of optimization programs from the literature are considered. In particular we revisit the scenario approach [7], [8], [9] and and the probabilistically robust design, a recently developed technique that is based on a mixture of randomized and robust optimization, proposed in [16]. Moreover, we consider the class of cascading optimization problems for which we provide novel results that offer guarantees regarding the probability of constraint satisfaction based on the compression learning perspective.

The rest of the paper unfolds as follows. Section 2 introduces the notion of compression. Section 3 shows how the learning theoretic results can be related to scenario based optimization and, in particular, the scenario approach and the probabilistically robust design. Section 4 shows how the proposed methodology can be employed for cascading optimization. Section 5 provides some discussion on the developed algorithms and elaborates on their relation with other learning based methodologies and Section 6 provides some concluding remarks. To simplify the presentation of the paper the proofs of each section have been moved to the corresponding appendix.

2. Learning results

2.1. Compression learning

We start by describing some concepts and results from compression learning introduced in [1]. We consider problems affected by an uncertain parameter δ taking values in some set $\Delta \subseteq \mathbb{R}^{n_\delta}$, which is endowed with a σ-algebra $\mathcal{D}$. Let $\mathbb{P}$ be a probability measure defined over $\mathcal{D}$. For $m \in \mathbb{N}$, we refer to a collection $\{\delta_i\}_{i=1}^m$ of m i.i.d. samples $\delta_i \in \Delta$ extracted according to $\mathbb{P}$ as an m-multisample.

We will refer to the elements $\mathcal{D}$ as *concepts*. For any concept $C \in \mathcal{D}$ let $\mathbb{1}_C(\cdot) : \Delta \to \{0,1\}$ be the standard indicator function of C, i.e. $\mathbb{1}_C(\delta) = 1$ if $\delta \in C$ and zero otherwise. Denote by $T \in \mathcal{D}$ a fixed but possibly unknown *target concept* for which we assume that an oracle is available, that for any $\delta \in \Delta$, provides the labeling $\mathbb{1}_T(\delta)$. The following basic definitions are adapted from [13].

Definition 1. *[*Labeled m-multisample*] Consider an m-multisample and a target concept $T \in \mathcal{D}$. A labeled m-multisample is the collection $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \in [\Delta \times \{0,1\}]^m$.*

Definition 2. *[*Consistent hypothesis*] Consider a labeled m-multisample and a target concept $T \in \mathcal{D}$. An element $H \in \mathcal{D}$ is called hypothesis. H is said to be consistent with the labeled m-multisample $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ if and only if $\mathbb{1}_H(\delta_i) = \mathbb{1}_T(\delta_i)$, for all $i = 1, \dots, m$.*

Definition 2 implies that H is a consistent hypothesis if it provides the same labeling of the samples δ_i, $i = 1, \dots, m$, as the target concept T. The error of H as an approximation of the target concept T can then be quantified through the probability measure of the set of uncertainty instances $\delta \in \Delta$ such that H and T give a different label. This error can be encoded by the measure of the symmetric difference of the sets T and H, i.e.

$$d_{\mathbb{P}}(T, H) = \mathbb{P}\big(\delta \in \Delta : \ \mathbb{1}_H(\delta) \neq \mathbb{1}_T(\delta)\big). \tag{1}$$

It is easy to see that $d_{\mathbb{P}}(T, H)$ is the measure[1] of the symmetric difference of the sets T and H. It is shown in [13] that $d_{\mathbb{P}}(\cdot, \cdot)$ is not a metric, but just a pseudo-metric, since $d_{\mathbb{P}}(C_1, C_2) = 0$ does not imply that $C_1 = C_2$, but only that the symmetric difference is a set of measure zero.

Definition 3. *[*Algorithm*] An algorithm is an indexed family of maps $\big\{A_m\big\}_{m \geq m_0}$ for some $m_0 \in \mathbb{N}$. The map $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ takes as input a labeled m-multisample and returns a hypothesis $A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$.*

The objective is to construct an approximation of the unknown target concept T by constructing an algorithm such that the hypothesis $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ is consistent with the m-multisample. Since H_m depends on the extracted multisample, it is a random quantity defined on the product space Δ^m with measure $\mathbb{P}^m$. We can therefore state the quality of the obtained approximation only probabilistically, determining the probability with respect to $\mathbb{P}^m$ with which the approximation error $d_{\mathbb{P}}(T, H_m)$ exceeds a given threshold.

Definition 4. *[PAC-T algorithm] Let $T \in \mathcal{D}$ be a target concept. Suppose there exists $m_0 \in \mathbb{N}$ so that the algorithm $\big\{A_m\big\}_{m \geq m_0}$ generates hypotheses $\big\{H_m\big\}_{m \geq m_0}$ such that for any $\epsilon \in (0,1)$, $m \geq m_0$,*

$$\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_m) > \epsilon\Big\} \leq q(m, \epsilon), \tag{2}$$

for some function $q(m, \epsilon) : \mathbb{N} \times (0,1) \to [0,1]$ such that $\lim_{m \to \infty} q(m, \epsilon) = 0$. Algorithm $\big\{A_m\big\}_{m \geq m_0}$ is then said to be Probably Approximately Correct for the target concept T (PAC-T).

[1] Throughout the paper we assume measurability of all involved sets. To relax this assumption the reader is referred to Appendix C in [17].

The statement of Definition 4 is clearly related to PAC learnability [13] (p. 56), where some concept class $\mathcal{C} \subseteq \mathcal{D}$ is considered and an algorithm is said to be PAC for the concept class $\mathcal{C}$ if (2) holds uniformly over target concepts $T \in \mathcal{C}$. Here we restrict attention to a specific target concept in view of the analysis of Section 3. For more details regarding PAC algorithms and PAC learnability the reader is referred to [13], [11].

Fix $d \in \mathbb{N}$ and consider $m \geq d$. We shall denote by $I_d = \{i_1, \ldots, i_d\}$ a set of d indices from $\{1, \ldots, m\}$ and by $\mathcal{I}_d$ the set of cardinality $\binom{m}{d}$ containing all I_d sets with d indices.

Theorem 1. *[Thm. 5 in [1]] Let $T \in \mathcal{D}$ be a target concept. Fix $d \in \mathbb{N}$, consider $m > d$ and denote by $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ a map that, for any $I_d \in \mathcal{I}_d$, takes as input the labeled d-multisample $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}$ and returns a hypothesis[2] $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}\big)$ consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}$. Then, for any $\epsilon \in (0,1)$ and any $m \geq d$*

$$\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \quad \text{there exists } I_d \in \mathcal{I}_d \text{ such that}$$
$$H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}. \tag{3}$$

Since for a fixed d, $\lim_{m\to\infty}\binom{m}{d}(1-\epsilon)^{m-d} = 0$, Theorem 1 implies that for a sufficiently high number of samples m, the probability that there exists a subset I_d with cardinality d of the m samples such that the hypothesis H_{I_d} generated by G_d is consistent with respect to all m samples but the approximation error exceeds ϵ is low. This theorem was stated in [1] in the context of sample compression, where the map G_d is referred to as the compression function.

Assumption 1. *Let $T \in \mathcal{D}$ be a target concept. Assume that there exists d and $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ taking as input a labeled d-multisample such that:*
1) For all $I_d \in \mathcal{I}_d$, H_{I_d} is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}$.
2) With $\mathbb{P}^m$-probability one, for any labeled m-multisample $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ with $m \geq d$, there exists $I_d \in \mathcal{I}_d$ such that the hypothesis $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}\big)$ is consistent with the labeled m-multisample.

Assumption 1 implies that any sufficiently large m-multisample can be compressed, i.e. there exists a subset of this multisample with fixed cardinality d which we can use to generate a hypothesis that is consistent with the entire m-multisample. The assumption that for any $I_d \in \mathcal{I}_d$, the hypothesis H_{I_d} is consistent with the d-multisample used to construct it, is trivially satisfied for the optimization problems considered in the next section.

Under Assumption 1, let the map $m_d : [\Delta \times \{0,1\}]^m \to \mathcal{I}_d$ return a set of d indices such that $G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in m_d}\big)$ is consistent with the entire $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Construct the algorithm $\big\{A_m\big\}_{m \geq d}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ takes as input a labeled m-multisample and returns a hypothesis

$$H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in m_d}\big). \tag{4}$$

We then have the following theorem, which is stated in [1] without a proof.

[2]Unlike H_m, the subscript of H_{I_d} is not an integer, but a set. The interpretation is that H_{I_d} is the output of the compression function when fed with the samples $\{\delta_i\}_{i \in I_d}$. In the sequel we use a similar notation when defining H_{m_d} for $m_d \in \mathcal{I}_d$. The reader is asked to excuse the slight abuse of the notation.

Theorem 2. *[Thm. 6 in [1]] Let $T \in \mathcal{D}$ be a target concept. Under Assumption 1, algorithm $\{A_m\}_{m\geq d}$ is PAC-T with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$.*

2.2. Strengthening the consistency assumption

Extending now the results of [1] we first show how the bound in Theorem 2 can be tightened by slightly strengthening Assumption 1.

Assumption 2. *Let $T \in \mathcal{D}$ be a target concept. Assume that there exists d and $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ taking as input a labeled d-multisample such that:*
1) For all $I_d \in \mathcal{I}_d$, H_{I_d} is consistent with $\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in I_d}$.
2) With $\mathbb{P}^m$-probability one, for any labeled m-multisample $\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i=1}^m$ with $m \geq d$, there exists a unique $I_d \in \mathcal{I}_d$ such that the hypothesis $H_{I_d} = G_d(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in I_d})$ is consistent with the labeled m-multisample.

The addition over Assumption 1 is that the set $I_d \in \mathcal{I}_d$ for which the requirements of Assumption 2 are satisfied is unique. For all $I_d \in \mathcal{I}_d$ define $S_{I_d} = \{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ H_{I_d}$ is consistent with $\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i=1}^m\}$, where $H_{I_d} = G_d(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in I_d})$. We then have the following propositions which are used in the proof of Theorem 3.

Proposition 1. *Under Assumption 2, $\{S_{I_d}\}_{I_d\in\mathcal{I}_d}$ forms a partition of Δ^m up to a set of measure zero, i.e. $\mathbb{P}^m\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ \Delta^m \setminus \cup_{I_d\in\mathcal{I}_d} S_{I_d}\} = 0$ and $S_{I_d^1} \cap S_{I_d^2} = \emptyset$ for all $I_d^1, I_d^2 \in \mathcal{I}_d$ with $I_d^1 \neq I_d^2$.*

Proposition 2. *Let $T \in \mathcal{D}$ be a target concept. Under Assumption 2, for any $I_d \in \mathcal{I}_d$ we have that*

$$F(\alpha) = \mathbb{P}^d\{\{\delta_i\}_{i\in I_d} \in \Delta^d : \ d_{\mathbb{P}}(T, H_{I_d}) \leq \alpha\} = \alpha^d, \tag{5}$$

where $F(\cdot)$ is the probability distribution of the error $d_{\mathbb{P}}(T, H_{I_d})$ and $\alpha \in [0,1]$.

The proof of Proposition 2 is similar to the first part of the proof of Theorem 1 in [8]. Define m_d, $\{A_m\}_{m\geq d}$ as in Section 2.1 and note that, under Assumption 2, $m_d : [\Delta \times \{0,1\}]^m \to \mathcal{I}_d$ is uniquely defined in this case.

Theorem 3. *Let $T \in \mathcal{D}$ be a target concept. Under Assumption 2, algorithm $\{A_m\}_{m\geq d}$ is PAC-T with $q(m,\epsilon) = \sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}$ and in particular, for any $\epsilon \in (0,1)$ and any $m \geq d$,*

$$\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_m) > \epsilon\Big\} = \sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}. \tag{6}$$

Theorem 3 constitutes a tighter version of Theorem 2 since (6) holds with equality for problems that satisfy Assumption 2. Moreover, the bound in the right-hand side of (6) is tighter compared to the one in Theorem 2. The proof of Theorem 3 is similar to the second part of the proof of Theorem 1 in [8].

2.3. Relaxing the consistency assumption

Finally, we revisit Theorem 2 and investigate relaxing Assumption 1. To this end fix $r,\ d \in \mathbb{N}$ and consider $m \geq d+r$. Given a set $I_r \in \mathcal{I}_r$, let the set $\mathcal{I}_d^{m-r}$ with cardinality $\binom{m-r}{d}$ contain all sets I_d with d indices from $\{1,\ldots,m\} \setminus I_r$.

Assumption 3. *Let $T \in \mathcal{D}$ be a target concept. Assume that there exists d and $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ taking as input a labeled d-multisample such that:*
1) For all $I_r \in \mathcal{I}_r$ and $I_d \in \mathcal{I}_d^{m-r}$, H_{I_d} is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}$.
2) With $\mathbb{P}^m$-probability one, for any labeled m-multisample $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ with $m \geq d+r$, for all $I_r \in \mathcal{I}_r$ there exists $I_d \in \mathcal{I}_d^{m-r}$ such that the hypothesis $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\ldots,m\}\setminus I_r}$,
3) With $\mathbb{P}^m$-probability one, for any labeled m-multisample $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ with $m \geq d+r$, there exists $I_r \in \mathcal{I}_r$ such that for any $I_d \in \mathcal{I}_d^{m-r}$ that satisfies the first part of the assumption, the hypothesis $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big)$ is not consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}$, for all $i \in I_r$.

The difference with Assumption 1 is that we now allow $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big)$ to be inconsistent with r elements of the labeled m-multisample. Suppose that Assumption 3 is satisfied and denote by $\bar{I}_r \in \mathcal{I}_r$ the set of indices such that the third part of the assumption holds. Let $\bar{m}_d^r : [\Delta \times \{0,1\}]^m \to \mathcal{I}_d$ be the map that for each labeled m-multisample $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ returns a set of d indices for which the corresponding hypothesis $G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\bar{m}_d^r}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\ldots,m\}\setminus\bar{I}_r}$ and is not consistent with $\big(\delta_i, \mathbb{1}_T(\delta_i)\big)$, for all $i \in \bar{I}_r$. Construct the algorithm $\big\{A_m\big\}_{m\geq d+r}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ takes as input a labeled m-multisample and returns a hypothesis $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\bar{m}_d^r}\big)$.

Theorem 4. *Let $T \in \mathcal{D}$ be a target concept and fix $r \in \mathbb{N}$. Under Assumption 3, algorithm $\big\{A_m\big\}_{m\geq d+r}$ is PAC-T with $q(m,\epsilon) = \binom{m}{d}\sum_{i=0}^r \binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i}$, i.e.*

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ d_{\mathbb{P}}(T, H_m) > \epsilon\Big\} \leq \binom{m}{d}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i}. \tag{7}$$

The proof of Theorem 4 is similar to the proof of Theorem 2.1 in [18]. We can strengthen Assumption 3 by requiring the set $I_d \in \mathcal{I}_d^{m-r}$ that satisfies its requirements to be unique. Consider now the following assumption, which is a relaxed version of Assumption 2.

Assumption 4. *Consider the set-up of Assumption 3. Assume also that the set $I_d \in \mathcal{I}_d^{m-r}$ that satisfies the requirements of Assumption 3 is unique.*

Consider the algorithm $\big\{A_m\big\}_{m\geq d+r}$, as constructed above Theorem 4. We then have the following theorem.

Theorem 5. *Let $T \in \mathcal{D}$ be a target concept and fix $r \in \mathbb{N}$. Under Assumption 4, algorithm $\big\{A_m\big\}_{m\geq d+r}$ is PAC-T with $q(m,\epsilon) = \binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}$, i.e.*

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ d_{\mathbb{P}}(T, H_m) > \epsilon\Big\} \leq \binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}. \tag{8}$$

The proof of Theorem 5 follows the proof of Theorem 2.1 in [18]. It constitutes a variant of Theorem 3 when Assumption 2 is relaxed to Assumption 4. However, in contrast to Theorem 3, the bound in (8) is not tight, since (63), (72) in the proof of Theorem 4 do not hold with equality.

3. Connection to optimization

3.1. Scenario based optimization as a learning problem

Consider the robust optimization problem

$$\begin{aligned}\mathcal{P}: &\min_{x\in\mathcal{X}} c^T x \\ &\text{subject to: } g(x,\delta)\le 0,\ \forall \delta\in\Delta,\end{aligned} \tag{9}$$

where $\mathcal{X}\subset\mathbb{R}^{n_x}$, $c\in\mathbb{R}^{n_x}$ and $g:\mathcal{X}\times\Delta\to\mathbb{R}$. As in Section 2 we assume that Δ is endowed with a σ-algebra and a probability measure $\mathbb{P}$. We consider here only one scalar-valued constraint function without loss of generality; in case of multiple constraint functions $g_j:\mathcal{X}\times\Delta\to\mathbb{R}$, $j=1,\dots,n_c$, we can set $g(x,\delta)=\max_{j=1,\dots,n_c} g_j(x,\delta)$. Moreover, considering a linear objective function is also without loss of generality; in case we seek to minimize a generic objective function, an epigraphic reformulation could be employed [7]. Optimization programs in the form of $\mathcal{P}$ are generally difficult to solve when Δ is a continuous set.

To determine an (approximate) solution to (9), an alternative optimization problem can be constructed, involving a multi-sample $\{\delta_i\}_{i=1}^m\in\Delta^m$ of finite size $m\in\mathbb{N}$, where the samples are extracted i.i.d according to $\mathbb{P}$.

$$\begin{aligned}\mathcal{P}[\{\delta_i\}_{i=1}^m]: &\min_{x\in\mathcal{X}} c^T x \\ &\text{subject to: } g(x,\delta)\le 0, \forall\delta\in S\big(\{\delta_i\}_{i=1}^m\big),\end{aligned} \tag{10}$$

where $S\big(\{\delta_i\}_{i=1}^m\big)\subseteq\Delta$ is a set that depends on the multisample; several choice of S will be presented in the sequel, among them $S\big(\{\delta_i\}_{i=1}^m\big)=\{\delta_i\}_{i=1}^m$.

In the set-up of Section 2, let $T=\Delta$ be the target concept, so that $\mathbb{1}_T(\delta)=1$ for all $\delta\in\Delta$. Fix $d\in\mathbb{N}$ and consider $m\ge d$ and any map $x_d:\Delta^d\to\mathcal{X}$. Define then a map $G_d:[\Delta\times\{0,1\}]^d\to\mathcal{D}$ such that for any $I_d\in\mathcal{I}_d$, it returns a hypothesis H_{I_d} constructed as

$$H_{I_d}=G_d\big(\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big)=\big\{\delta\in\Delta:\ g(x_d(\{\delta_i\}_{i\in I_d}),\delta)\le 0\big\}. \tag{11}$$

Since $T=\Delta$, for any $I_d\in\mathcal{I}_d$, $d_{\mathbb{P}}(T,H_{I_d})$ is the probability of constraint violation, i.e.

$$d_{\mathbb{P}}(T,H_{I_d})=\mathbb{P}(\{\delta\in\Delta:\ \delta\notin H_{I_d}\})=\mathbb{P}\big(\{\delta\in\Delta:\ g(x_d(\{\delta_i\}_{i\in I_d}),\delta)>0\}\big). \tag{12}$$

Suppose that d, G_d are such that Assumption 1 is satisfied. Then there exists $m_d(\delta_1,\dots,\delta_m)\in\mathcal{I}_d$ such that $H_{m_d}=\big\{\delta\in\Delta:\ g(x_d(\{\delta_i\}_{i\in m_d}),\delta)\le 0\big\}$ is consistent with $\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Note that Assumption 1 implicitly requires H_{m_d} to be non-empty, since it must include $\{\delta_i\}_{i=1}^m$. This implies that $x_d(\{\delta_i\}_{i\in m_d})$ is feasible for $\mathcal{P}[\{\delta_i\}_{i=1}^m]$.

Theorem 6. *Let $T=\Delta$ be the target concept and consider Assumption 1. Let $x_m:\Delta^m\to\mathcal{X}$ be such that $x_m(\{\delta_i\}_{i=1}^m)=x_d(\{\delta_i\}_{i\in m_d})$ for a set $m_d\in\mathcal{I}_d$ that satisfies the second part of Assumption 1. Then, for any $\epsilon\in(0,1)$,*

$$\mathbb{P}^m\Big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ \mathbb{P}\big(\delta\in\Delta:\ g(x_m(\{\delta_i\}_{i=1}^m),\delta)>0\big)>\epsilon\Big\}\le\binom{m}{d}(1-\epsilon)^{m-d}. \tag{13}$$

Theorem 6 shows that under Assumption 1, for any feasible solution x_m of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ such that $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$, we can provide probabilistic guarantees regarding its feasibility of the form of (13). Note that the statement of (13) shows that, with probability at least $1-\binom{m}{d}(1-\epsilon)^{m-d}$, x_m satisfies (9) except for a set with $\mathbb{P}$-measure at most ϵ. The proof of Theorem 6 is based on showing that an algorithm is PAC-T for the target concept $T = \Delta$. This algorithm can be constructed as $\{A_m\}_{m\geq d}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ is such that $H_m = A_m\big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i=1}^m\big)$ and $H_m = H_{m_d}$. The hypothesis H_m is defined as $H_m = \{\delta \in \Delta : \ g(x_m(\{\delta_i\}_{i=1}^m), \delta) \leq 0\}$, while ensuring that $H_m = H_{m_d}$ is equivalent to $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$. The latter is satisfied in the scenario approach set-up of Section 3.2 and the probabilistically robust design of Section 3.3.

Note that if we replace Assumption 1 with Assumption 2, Theorem 6 is still valid with the right-hand side of (13) being replaced by the right-hand side of (6) in Theorem 3; in fact the result would hold with equality. Following the discussion at the end of Section 2.3, one could also relax Assumption 1 in a way such that the right-hand side of (13) is replaced by $\binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}$. The interpretation of a hypothesis that is not consistent with some elements of the multi-sample in an optimization context is that we allow for some of the constraints to be violated. For problems that are convex with respect to the decision variables, this procedure is referred to as sampling-and-discarding in [18] and as constraint removal in [9].

We next consider problems for which probabilistic feasibility guarantees similar to (13) are provided in [7], [16], following a different methodology. Here we adopt the compression learning perspective and show that these problems share certain similarities, thus justifying the fact their guarantees are of the same form. In particular, we show that by appropriately selecting the constraint function, the uncertainty set of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and the map $x_m : \Delta^m \to \mathcal{X}$, the requirements of Assumption 1 are satisfied, and hence we obtain the probabilistic feasibility guarantees by virtue of Theorem 6.

3.2. The scenario approach

We first present the set-up of the scenario approach as this was proposed in [7]. For any $m \in \mathbb{N}$ consider $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ with $S(\{\delta_i\}_{i=1}^m) = \{\delta_i\}_{i=1}^m$; this results in the following optimization problem.

$$\begin{aligned}\mathcal{P}_1[\{\delta_i\}_{i=1}^m] : \min_{x\in\mathcal{X}} \ & c^T x \\ \text{subject to: } & g(x,\delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m,\end{aligned} \tag{14}$$

Let $\mathcal{X}_m = \{x \in \mathcal{X} : \ g(x,\delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\}$ be the feasibility region of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ and consider the following assumption.

Assumption 5. *The set $\mathcal{X} \subset \mathbb{R}^{n_x}$ is convex and for any $\delta \in \Delta$, the constraint function $g(\cdot,\delta)$ is convex. For any m-multisample $\{\delta_i\}_{i=1}^m$, the feasibility region $\mathcal{X}_m$ of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ has a non-empty interior and the minimizer of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ exists and is unique.*

The uniqueness and the feasibility part of the assumption can be relaxed as shown in [8], [9]. However, we keep these assumptions here to simplify the presentation. Under Assumption 5, let x_m be the minimizer of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ and note that x_m belongs to the feasibility region of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$.

The scenario approach is based on the notion of support constraints. A constraint in $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ is said to be a support constraint, if its removal results in an improvement in the objective value (see also Definition 4 in [7]). In [9], under the convexity part of Assumption 5, it is shown that, with $\mathbb{P}^m$-probability one, the number of support constraints is bounded by the so called Helly's dimension.

In [7], [8] it is shown that Helly's dimension is upper-bounded by n_x, whereas in [19] an improved bound is provided. Let the number of support constraints be *at most* $\zeta < \infty$. Under Assumption 5, and based on the definition of the support constraints, it can be shown that Assumption 1 is satisfied for $d = \zeta$ and an appropriately constructed map G_d.

Proposition 3. *Let $T = \Delta$ be the target concept and consider Assumption 5. Fix $d = \zeta$ and consider $m \geq d$. For any $I_d \in \mathcal{I}_d$, let $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ return a hypothesis $H_{I_d} = \big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in I_d}), \delta) \leq 0\big\}$, where x_d is the minimizer of $\mathcal{P}_1[\{\delta_i\}_{i\in I_d}]$. G_d then satisfies Assumption 1.*

Under Proposition 3, there exists $m_d \in \mathcal{I}_d$ with $d = \zeta$ such that the hypothesis $H_{m_d} = \big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in m_d}), \delta) \leq 0\big\}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Moreover, as shown in the proof of Proposition 3, the set m_d for which Assumption 1 is satisfied is such that $x_d\big(\{\delta_i\}_{i\in m_d}\big) = x_m\big(\{\delta_i\}_{i=1}^m\big)$, where x_m is the unique (under Assumption 5) minimizer of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$. This leads to the following corollary of Theorem 6.

Corollary 1. *Let $T = \Delta$ be the target concept and consider Assumption 5. Fix $d = \zeta$ and consider $m \geq d$. Then, for any $\epsilon \in (0,1)$,*

$$\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m :\ \mathbb{P}\big(\delta \in \Delta :\ g(x_m, \delta) > 0\big) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}, \tag{15}$$

where x_m is the minimizer of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$.

Corollary 1 provides guarantees on the probability that the optimal solution of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ violates the constraints. Note that this result is identical to Theorem 1 of [7] (with n_x in place of ζ) but is not the same with the refined bound of Theorem 1 of [8]. To obtain the same conclusion with Theorem 1 of [8] we focus first on problems in the form of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ such that, with $\mathbb{P}^m$-probability one, the number of support constraints is *equal to* ζ. In the particular case where $d = \zeta = n_x$, we have the class of fully supported problems [8]. Considering problems where $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ has exactly ζ support constraints with probability one, is a sufficient condition for Assumption 2 to be satisfied. This is summarized in the following proposition.

Proposition 4. *Let $T = \Delta$ be the target concept and consider Assumption 5. Fix $d = \zeta$ and consider $m \geq d$. For any $I_d \in \mathcal{I}_d$, let $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ return a hypothesis $H_{I_d} = \big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in I_d}), \delta) \leq 0\big\}$, where x_d is the minimizer of $\mathcal{P}_1[\{\delta_i\}_{i\in I_d}]$. If $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ has exactly ζ support constraints with $\mathbb{P}^m$-probability one, then G_d satisfies Assumption 2.*

We then have the following corollary.

Corollary 2. *Let $T = \Delta$ be the target concept and consider Assumption 5. Suppose that $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ has exactly ζ support constraints with $\mathbb{P}^m$-probability one. Fix $d = \zeta$ and consider $m \geq d$. Then, for any $\epsilon \in (0,1)$,*

$$\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m :\ \mathbb{P}\big(\delta \in \Delta :\ g(x_m, \delta) > 0\big) > \epsilon\Big\} = \sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}, \tag{16}$$

where x_m is the minimizer of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$.

If the problem does not have exactly ζ support constraints with $\mathbb{P}^m$-probability one, we can still obtain similar probabilistic guarantees following [9], [8]. Specifically, it is shown that if a problem is non-degenerate (see [9] for a definition of non-degenerate problems) and has at most ζ support constraints, then by a procedure called regularization it can be transformed to a different problem with exactly ζ support constraints. We can then bound the probability in the left-hand side of (13) by the probability of constraint violation for the regularized problem, which is equal to $\sum_{i=0}^{\zeta-1} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i}$. In [8] it is shown that this is also the case even for degenerate problems that do not have exactly ζ support constraints.

We can replace Assumption 1 in Proposition 3 and Assumption 2 in Proposition 4 by Assumption 3 and Assumption 4, respectively. The right-hand side of (15) is then replaced by the right-hand side of (7). Similarly, the right-hand side of (16) is replaced by the right-hand side of (8), but the result does not necessarily hold with equality. However, note that Assumption 5 does not suffice to ensure that both parts of Assumption 3 (similarly for Assumption 4) are satisfied; it only guarantees (via Proposition 3 with $m-r$ in place of m) that the requirement of the first part holds. To ensure that requirement of the second part is also satisfied we equip the algorithm constructed in the proof of Propositions 3 and 4 by a procedure that removes r samples such that the minimizer of the problem with the remaining $m-r$ samples violates all constraints that correspond to the removed samples. As an effect of this removal procedure the objective value is always decreasing every time a sample is removed.

Such a procedure is referred to as sampling-and-discarding in [18] and as scenario approach with constraint removal in [9]. Moreover, in [18], [9], different methodologies to construct such a procedure are proposed and their complexity is discussed: an optimal constraint removal scheme, however, with a combinatorial complexity; a greedy approach where the r constraints to be removed are eliminated on a sequential fashion; and an approach based on the Lagrange multipliers associated with the constraint functions.

3.3. Probabilistically robust design

We now revisit the probabilistically robust design proposed in [16]. For any $m \in \mathbb{N}$ consider the following optimization problem:

$$\begin{aligned} \widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m] : \min_{\underline{p}, \overline{p} \in \mathbb{R}^{n_\delta}} \; & ||\overline{p} - \underline{p}||_1 \\ \text{subject to: } & \delta \in \big[\underline{p},\ \overline{p}\big], \forall \delta \in \{\delta_i\}_{i=1}^m, \end{aligned} \tag{17}$$

where the inclusion in (17) should be interpreted element-wise. Denote by $p_m = (\underline{p}_m,\ \overline{p}_m) \in \mathbb{R}^{2n_\delta}$ the minimizer of $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m]$, which depends on the multisample $\{\delta_i\}_{i=1}^m$. Let $B(p_m) \subset \Delta$ be a hyper-rectangle constructed by the cartesian product of the intervals in $\big[\underline{p}_m,\ \overline{p}_m\big]$. Clearly, $B(p_m)$ is the smallest axis-aligned hyper-rectangle that contains all samples $\{\delta_i\}_{i=1}^m$. Consider now the following optimization problem:

$$\begin{aligned} \mathcal{P}_2[\{\delta_i\}_{i=1}^m] : \min_{x \in \mathcal{X}} \; & c^T x \\ \text{subject to: } & g(x, \delta) \leq 0, \forall \delta \in B(p_m). \end{aligned} \tag{18}$$

Problem $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ is a robust program and requires the constraints to be satisfied for all values of the uncertainty inside $B(p_m)$, which is constructed based on the optimal solution of $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m]$. Note that $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ is of the same form with $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ with $S(\{\delta_i\}_{i=1}^m) = B\big(p_m(\{\delta_i\}_{i=1}^m)\big)$. For a

detailed discussion regarding conditions under which $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ is tractable, the reader is referred to [16].

Let now $\mathcal{X}_m = \{x \in \mathcal{X} : \ g(x,\delta) \leq 0, \forall \delta \in B(p_m)\}$ be the feasibility region of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ and consider the following assumption.

Assumption 6. *For any m-multisample $\{\delta_i\}_{i=1}^m$, the feasibility region $\mathcal{X}_m$ of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ has a non-empty interior, and the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ exists and is unique.*

Under Assumption 6, let x_m to be the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$. Note that for any $\{\delta_i\}_{i=1}^m$, $x_m(\{\delta_i\}_{i=1}^m) \in \mathcal{X}_m$. Imposing the uniqueness assumption and selecting x_m to be the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ is to simplify the presentation of our results and at the end of the section we remove the uniqueness part of the assumption and discuss alternative choices for the map x_m.

Proposition 5. *Let $T = \Delta$ be the target concept and consider Assumption 6. Fix $d = 2n_\delta$ and consider $m \geq d$. For any $I_d \in \mathcal{I}_d$, let $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ return a hypothesis $H_{I_d} = \big\{\delta \in \Delta : \ g(x_d(\{\delta_i\}_{i\in I_d}),\delta) \leq 0\big\}$, where x_d is the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i\in I_d}]$. G_d then satisfies Assumption 1.*

Under Proposition 3, there exists $m_d \in \mathcal{I}_d$ with $d = 2n_\delta$ such that the hypothesis $H_{m_d} = \big\{\delta \in \Delta : \ g(x_d(\{\delta_i\}_{i\in m_d}),\delta) \leq 0\big\}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. In fact, as shown in the proof of Proposition 5, there exists a unique set of indices $m_d \in \mathcal{I}_d$ satisfying Assumption 1. Moreover, this set is such that $B(p_d(\{\delta_i\}_{i\in m_d})) = B(p_m(\{\delta_i\}_{i=1}^m))$, where p_d is the minimizer of $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i\in m_d}]$. The latter implies that $\mathcal{X}_{m_d} = \mathcal{X}_m$, where $\mathcal{X}_{m_d}$ is the feasibility region of $\mathcal{P}_2[\{\delta_i\}_{i\in m_d}]$. Due to the uniqueness part of Assumption 6 we then have that $x_d(\{\delta_i\}_{i\in m_d}) = x_m(\{\delta_i\}_{i=1}^m)$, where x_m is the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$. This leads to the following corollary of Theorem 6.

Corollary 3. *Let $T = \Delta$ be the target concept and consider Assumption 6. Fix $d = 2n_\delta$ and consider $m \geq d$. Then, for any $\epsilon \in (0,1)$,*

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m : \ \mathbb{P}\big(\delta \in \Delta : \ g(x_m,\delta) > 0\big) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}, \tag{19}$$

where x_m is the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$.

In general we can provide guarantees in the form of (19) for any a-priori specified map x_m that determines some feasible solution of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$, and not only for the minimizer. Remove now the uniqueness requirement of Assumption 6. We show that, by selecting x_d according to the following procedure, we obtain guarantees for the entire feasibility region $\mathcal{X}_m$ of the robust problem $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$. To achieve this, for any $I_d \in \mathcal{I}_d$, consider the worst case probability of constraint violation $\sup_{x\in\mathcal{X}_{I_d}} \mathbb{P}\big(\delta \in \Delta : \ g(x,\delta) > 0\big)$. Then, for any $\bar{\epsilon} > 0$ there exists $x_d[\bar{\epsilon}] : \Delta^d \to \mathcal{X}$ with $x_d[\bar{\epsilon}](\{\delta_i\}_{i\in I_d}) \in \mathcal{X}_{I_d}$ such that

$$\sup_{x\in\mathcal{X}_{I_d}} \mathbb{P}\big(\delta \in \Delta : \ g(x,\delta) > 0\big) < \mathbb{P}\big(\delta \in \Delta : \ g(x_d[\bar{\epsilon}],\delta) > 0\big) + \bar{\epsilon}. \tag{20}$$

For $\bar{\epsilon} > 0$ pick any such $x_d[\bar{\epsilon}]$. Under this choice it can be shown that (19) can be replaced by

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m : \ \sup_{x\in\mathcal{X}_m} \mathbb{P}\big(\delta \in \Delta : \ g(x,\delta) > 0\big) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}. \tag{21}$$

The proof of this statement is similar to the proof of Corollary 3 and relies on the fact that for the set $m_d(\{\delta_i\}_{i=1}^m)$ of indices satisfying Assumption 1, for any $\bar{\epsilon} > 0$, $x_d[\bar{\epsilon}]$ satisfies (19) and, as shown in the proof of Proposition 5, $\mathcal{X}_{m_d} = \mathcal{X}_m$. Equation (21) follows then from (20) and the fact that $\bar{\epsilon} > 0$ is arbitrary.

The result in (21) is similar but not identical to Proposition 1 of [16] where a tighter bound is provided; however, we achieve these guarantees by means of Theorem 6 without resorting to the scenario approach as in [16]. The rest of the section demonstrates how we can obtain the same conclusion with Proposition 1 of [16]. To this end consider the following proposition.

Proposition 6. *Let $T = \Delta$ be the target concept and consider Assumption 6. Fix $d = 2n_\delta$ and consider $m \geq d$. For any $I_d \in \mathcal{I}_d$, let $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ return a hypothesis $H_{I_d} = \big\{\delta \in \Delta : \ g(x_d(\{\delta_i\}_{i\in I_d}), \delta) \leq 0\big\}$, where x_d is the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i\in I_d}]$. If, with $\mathbb{P}^m$-probability one, for any $I_d \in \mathcal{I}_d$*

$$\Big\{\delta \in \Delta : \ g(x_d(\{\delta_i\}_{i\in I_d}), \delta) > 0\Big\} = \Big\{\delta \in \Delta : \ \delta \notin B(p_d(\{\delta_i\}_{i\in I_d}))\Big\}, \tag{22}$$

then G_d satisfies Assumption 2.

We then have the following corollary.

Corollary 4. *Let $T = \Delta$ be the target concept and consider Assumption 6. Fix $d = \zeta$ and consider $m \geq d$. Suppose that with $\mathbb{P}^m$-probability one, equality (22) is also satisfied for any $I_d \in \mathcal{I}_d$. Then, for any $\epsilon \in (0,1)$,*

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m : \ \mathbb{P}\big(\delta \in \Delta : \ g(x_m,\delta) > 0\big) > \epsilon\Big\} = \sum_{i=0}^{d-1} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i}, \tag{23}$$

where x_m is the minimizer of $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$.

If (22) is not satisfied, Corollary 4 does not hold any more; this is not the case with Corollary 3. However, by inspection of (18) we have that, for any $x \in \mathcal{X}_m$, if $\delta \in B(p_m)$ then $g(x,\delta) \leq 0$. Since the last statement holds for any $x \in \mathcal{X}_m$ it will also hold for x_m. Therefore, $\mathbb{P}\big(\delta \in \Delta : \ g(x_m,\delta) > 0\big) \leq \mathbb{P}\big(\delta \in \Delta : \ \delta \notin B(p_m)\big)$, and hence $\mathbb{P}^m\big\{(\delta_1,\ldots,\delta_m) \in \Delta^m \ : \mathbb{P}\big(\delta \in \Delta : \ g(x_m,\delta) > 0\big) > \epsilon\big\} \leq \mathbb{P}^m\big\{(\delta_1,\ldots,\delta_m) \in \Delta^m \ : \mathbb{P}\big(\delta \in \Delta : \ \delta \notin B(p_m)\big) > \epsilon\big\}$. The right-hand side of the previous inequality corresponds to the probability with respect to $\mathbb{P}^m$ that the probability of constraint violation of $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m]$ exceeds ϵ. The latter falls in the framework of the scenario approach and has by construction $\zeta = 2n_\delta$ support constraints. In fact, $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=m}^m]$ is a fully-supported problem with $2n_\delta$ decision variables. Therefore, for any $\epsilon > 0$, Assumption 2 is satisfied for this problem and Corollary 2 holds with $d = \zeta = 2n_\delta$. Hence,

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m : \ \mathbb{P}\big(\delta \in \Delta \mid \delta \notin B(p_m)\big) > \epsilon\Big\} = \sum_{i=0}^{2n_\delta-1} \binom{m}{i} \epsilon^i \big(1-\epsilon\big)^{m-i}. \tag{24}$$

Therefore, in any case we have that

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m : \ \mathbb{P}\big(\delta \in \Delta : \ g(x_m,\delta) > 0\big) > \epsilon\Big\} \leq \sum_{i=0}^{2n_\delta-1} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i}. \tag{25}$$

Note, however, that (25) is not tight. Moreover, selecting the map x_d as in (20) we can provide guarantees for the entire feasibility region $\mathcal{X}_m$, and replace the left-hand side in (25) by $\mathbb{P}^m\Big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ \sup_{x\in\mathcal{X}_m}\mathbb{P}\big(\delta\in\Delta:\ g(x,\delta)>0\big)>\epsilon\Big\}$. However, due to (20), the inequality in (25) would be strict.

We can replace Assumption 1 in Proposition 5 and Assumption 2 in Proposition 6 to Assumption 3 and Assumption 4, respectively. The right-hand side of (19) is then replaced by the right-hand side of (7). Similarly, the right-hand side of (23) is replaced by the right-hand side of (8), but the result does not necessarily hold with equality. However, note that Assumption 6 does not suffice to ensure that both parts of Assumption 3 (similarly for Assumption 4) are satisfied; it only guarantees (via Proposition 3 with $m-r$ in place of m) that the requirement of the first part holds. Similarly to the scenario approach set-up, to ensure that requirement of the second part is also satisfied we equip the algorithm constructed in the proof of Propositions 5 and 4 by a procedure that removes r samples such that the minimizer of the problem with the remaining $m-r$ samples violates all constraints that correspond to the removed samples. As an effect of this removal procedure the objective value is always decreasing every time a sample is removed.

In [16] one such procedure is proposed. First $\mathcal{P}_2[\{\delta_i\}_{i=1}^m]$ is solved and the samples that correspond to the active constraints of $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m]$ are identified. In fact these samples are the ones that lie on the facets of $B(p_m)$. From these samples remove δ_j, for some $j\in\{1,\dots,m\}$ that yields the highest reduction in the objective value of $\mathcal{P}_2[\{\delta_i\}_{i\in\{1,\dots,m\}\setminus j}]$ (this implies that the feasibility region is enlarged). Typically, this step requires solving $2n_\delta$ (assuming no multiple samples on the same facet of $B(p_m)$) robust optimization problems. We then proceed the same way until r samples are removed. Similarly to the scenario approach, as an effect of this removal procedure the objective value of the robust problem is always decreasing every time a sample is removed.

Note that for $m\in\mathbb{N}$, we selected $B(p_m)$ to be a hyper-rectangle. However, any other representation (e.g. sphere, polytope, ellipsoid) with fixed parametrization could have been chosen instead, by reformulating $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m]$ as a convex volume minimization problem. In that case our analysis would remain unchanged with $2n_\delta$ being replaced by the dimension of the parametrization vector p_m. For example, if $B(p_m)$ is a sphere, we would need $n_\delta+1$ parameters.

4. Cascading optimization problems

4.1. Probabilistic performance guarantees

We consider here the class of cascading optimization problems. Every problem in the cascade is a program that depends on uncertainty scenarios but also on the solution of the preceding problem, while the same uncertainty scenarios are used in all problems in the cascade. Such problems arise in different contexts (e.g. multi-objective optimization, bilinear descent type of algorithms, approximate dynamic programming), yet, to the best of our knowledge, obtaining guarantees regarding the probability of constraint violation for the solution comprising the solutions of the individual problems in the cascade has proven to be elusive. Our analysis provides such guarantees for a cascade of two problems, but our results can be immediately extended to the case of any finite number of cascading problems.

For any $m\in\mathbb{N}$, consider the following family of problems which is parametric in $x\in\mathcal{X}$:

$$\begin{aligned}\widetilde{\mathcal{P}}[x,\{\delta_i\}_{i=1}^m]:\min_{y\in\mathcal{Y}}\ &\tilde{c}^T y\\ &\text{subject to: } \widetilde{g}(y,x,\delta)\le 0,\forall\delta\in\widetilde{S}\big(\{\delta_i\}_{i=1}^m\big),\end{aligned}\tag{26}$$

where $x \in \mathcal{X}$ is the vector of decision variables of an optimization problem of the form of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ in (10), $\mathcal{Y} \subset \mathbb{R}^{n_y}$, $\tilde{c} \in \mathbb{R}^{n_y}$, $\widetilde{g}:\ \mathcal{Y} \times \mathcal{X} \times \Delta \to \mathbb{R}$, and $\widetilde{S}\big(\{\delta_i\}_{i=1}^m\big) \subseteq \Delta$.

Suppose that $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i=1}^m]$, for all $x \in \mathcal{X}$, fall in the scenario approach set-up, i.e. $\widetilde{S}\big(\{\delta_i\}_{i=1}^m\big) = S\big(\{\delta_i\}_{i=1}^m\big) = \{\delta_i\}_{i=1}^m$. Therefore, we impose the following assumption.

Assumption 7. *Suppose that $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ is in the form of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ in (14) satisfying Assumption 5. Moreover, $\widetilde{S}\big(\{\delta_i\}_{i=1}^m\big) = \{\delta_i\}_{i=1}^m$, the set $\mathcal{Y} \subset \mathbb{R}^{n_y}$ is convex and for any $x \in \mathcal{X}$ and any $\delta \in \Delta$, the constraint function $\widetilde{g}(\cdot, x, \delta)$ is convex. For any $x \in \mathcal{X}$ and any m-multisample $\{\delta_i\}_{i=1}^m$, the feasibility region $\big\{y \in \mathcal{Y}:\ \widetilde{g}(y, x, \delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\}$ of $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i=1}^m]$ has a non-empty interior and the minimizer of $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i=1}^m]$ exists and is unique.*

We only need to invoke Assumption 7 in the proof of Proposition 7 and Theorem 7, where a by-product of Proposition 3 is employed. Alternatively, we could assume that problems $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i=1}^m]$, for all $x \in \mathcal{X}$, fall in the set-up of the probabilistically robust design and modify Assumption 7 so that both problems satisfy Assumption 6. Moreover, even if these problems belong to any problem class, the subsequent developments would still follow, as long as the solution of each problem does not alter if we only use the subset of the m-multisample returned by the compression function defined below.

Under Assumption 7, Proposition 3 implies that Assumption 1 is satisfied for some $d_1 \in \mathbb{N}$, $G_{d_1} : [\Delta \times \{0,1\}]^{d_1} \to \mathcal{D}$, which for any $I_{d_1} \in \mathcal{I}_{d_1}$ returns $H_{I_{d_1}} = \big\{\delta \in \Delta:\ g(x_{d_1}(\{\delta_i\}_{i \in I_{d_1}}), \delta) \leq 0\big\}$, where $x_{d_1}:\ \Delta^{d_1} \to \mathcal{X}$ is the minimizer of $\mathcal{P}[\{\delta_i\}_{i \in I_{d_1}}]$. Similarly, for any $x \in \mathcal{X}$, Assumption 1 is also satisfied for some $d_2 \leq n_y$, $\widetilde{G}_{d_2}[x] : [\Delta \times \{0,1\}]^{d_2} \to \mathcal{D}$, which for any $I_{d_2} \in \mathcal{I}_{d_2}$ returns the hypothesis $\widetilde{H}_{I_{d_2}}[x] = \widetilde{G}_{d_2}[x]\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_{d_2}}\big) = \big\{\delta \in \Delta:\ \widetilde{g}(y_{d_2}[x](\{\delta_i\}_{i \in I_{d_2}}), x, \delta) \leq 0\big\}$, where $y_{d_2}[x] : \Delta^{d_2} \to \mathcal{Y}$ is the unique, under Assumption 7, minimizer of $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i \in I_{d_2}}]$.

Proposition 7. *Let $T = \Delta$ be the target concept and consider Assumption 5. Fix $d = d_1 + d_2$ and consider $m \geq d$. Construct $G_d^c : [\Delta \times \{0,1\}]^d \to \mathcal{D}$, such that for any $I_d \in \mathcal{I}_d$,*

$$\begin{aligned} G_d^c\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}\big) &= H_{I_d} \cap \widetilde{H}_{I_d}[x_d(\{\delta_i\}_{i \in I_d})] \\ &= \big\{\delta \in \Delta:\ \big(g(x_d(\{\delta_i\}_{i \in I_d}), \delta) \leq 0\big) \text{ and } \big(\widetilde{g}(y_d[x_d(\{\delta_i\}_{i \in I_d})](\{\delta_i\}_{i \in I_d}), x_d(\{\delta_i\}_{i \in I_d}), \delta) \leq 0\big)\big\}. \end{aligned} \tag{27}$$

G_d^c then satisfies Assumption 1.

Proposition 7 shows that if there exist a compression function for two optimization problems, then there exists a compression function for the cascade of these problems, where the outcome of the latter depends on the solution of the former. Under Proposition 7, there exists $m_d\big(\{\delta_i\}_{i=1}^m\big) \in \mathcal{I}_d$ such that the hypothesis $H_{m_d}^c = G_d^c\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in m_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$.

Theorem 7. *Let $T = \Delta$ be the target concept and consider Assumption 7. Fix $d = d_1 + d_2$ and consider $m \geq d$. Then, for any $\epsilon \in (0,1)$,*

$$\begin{aligned} &\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m: \\ &\mathbb{P}\Big(\delta \in \Delta:\ \big(g(x_m, \delta) > 0\big) \text{ or } \big(\widetilde{g}(y_m[x_m], x_m, \delta) > 0\big)\Big) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}, \end{aligned} \tag{28}$$

where x_m and $y_m[x_m]$ are the minimizers of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and $\widetilde{\mathcal{P}}[x_m, \{\delta_i\}_{i=1}^m]$, respectively.

Theorem 7 provides a bound on the probability with which x_m, y_m violate either the constraints of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$, or the constraints of $\widetilde{\mathcal{P}}[x_m, \{\delta_i\}_{i=1}^m]$. Its proof is based on showing that an algorithm, $\{A_m\}_{m\geq d}$, is PAC-T for the target concept $T = \Delta$. This algorithm comprises $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ such that $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ and $H_m = H_{m_d}^c$. The hypothesis H_m is defined as $H_m = \big\{\delta \in \Delta :\ \big(g(x_m,\delta) \leq 0\big)$ or $\big(\widetilde{g}(y_m[x_m], x_m, \delta) \leq 0\big)\big\}$. Ensuring that $H_m = H_{m_d}$ is equivalent to $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$ and $y_m[x_m](\{\delta_i\}_{i=1}^m) = y_d[x_d](\{\delta_i\}_{i\in m_d})$. The latter follows from the proof of Proposition 3. We refer to $\{A_m\}_{m\geq d}$ as cascading algorithm since it is constructed based on a cascade of two sequentially dependent hypotheses.

Note that, under Assumption 7, we need $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i=1}^m]$ to be feasible for any $x \in \mathcal{X}$. To relax this requirement consider the set

$$F = \Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m :\ \forall x \in \big\{x \in \mathcal{X} :\ g(x,\delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\},$$
$$\big\{y \in \mathcal{Y} :\ \widetilde{g}(y, x, \delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\} \neq \emptyset\Big\}. \tag{29}$$

F is a restriction of Δ^m on the set of multisamples for which the second problem in the cascade has a non-empty feasibility region (feasibility of the first one is ensured under Assumption 5), not for any $x \in \mathcal{X}$, but for any $x \in \big\{x \in \mathcal{X} :\ g(x,\delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\}$. The result of Theorem 7 will then still hold if we replace Δ^m with F in (28).

Theorem 7 implies that the solution comprising the solutions of the individual problems in the cascade is feasible for the constraints of both problems. In certain cases one can obtain similar guarantees by formulating a single optimization problem that involves minimizing some convex objective function (e.g. the objective function of the last problem in the cascade) with respect to both $x \in \mathbb{R}^{n_x}$ and $y \in \mathbb{R}^{n_y}$, and subject to the constraints of both problems in the cascade. However, guarantees in the form of (28) can be still provided only if the second problem in the cascade is jointly convex with respect to x and y. This is not required with the proposed approach, and the second problem in the cascade is allowed to have an arbitrary dependence with respect to x (see Assumption 7). Moreover, even if the constraint functions are convex with respect to the decision variables of both problems, solving a single program involving all constraints may result to solutions x, y that are not optimal for the individual problems in the cascade, thus leading to a degraded objective value. One example of a problem with constraint functions that are not jointly convex with respect to the decision variables x and y can be found in bilinear descent type of algorithms. Suppose we seek to minimize some convex objective function subject to constraints that should hold for all $\delta \in \{\delta_i\}_{i=1}^m$, and the constraint functions are bi-convex with respect to x and y. One way to deal with this problem is by applying an iterative procedure with an a-priori fixed number of iterations. We could arbitrarily fix $y = y_0$ and consider the problem of minimizing only with respect to x. The resulting problem would then be in the form of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$. Let x_m be the minimizer of this problem. We can then fix $x = x_m$ in the initial problem and minimize only with respect to y. If we do not follow such an iterative approach, since the problem is non-convex, to provide guarantees in the form of (28) one should resort to VC theory, which involves, however, the computation of an upper bound of the VC dimension, which is not necessarily easy to determine.

Another important feature of the proposed approach is that in both $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i=1}^m]$ the same samples $\{\delta_i\}_{i=1}^m$ are used. This is required, for example, in the stochastic model predictive control context considered in [20], where a cascade of two scenario programs was formulated. The first problem in the cascade was in the form of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ with the constraint function encoding the input constraints (depending on samples). At the second problem in the cascade, the bound on the system sate was considered as a decision variable. The objective was to minimize this (soft) bound,

subject to both input and state constraints (depending on the same samples with the first problem) and the additional constraint $c^T y \leq c^T x_m + \alpha$, where x_m is the minimizer of the first problem, y includes the decision variables of the second problem and $\alpha > 0$ is a pre-specified degradation parameter. The second problem is then also in the form of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$. This two-step approach has a multi-objective nature since it allows us to relax the state constraints by deciding upon their bound in the second problem in the cascade, while ensuring that the objective value deteriorates at most by a fixed amount α compared to the value obtained at the first problem. In particular, the two problems in the cascade have the same decision variables, i.e. $x = y$ and $n_x = n_y$, and the set F in (29) is such that $F = \Delta^m$. The same samples have to be employed in both problems, otherwise feasibility of the second problem is not guaranteed. This is also the case in bilinear descent type of algorithms, since by using the same samples at every problem in the cascade, the objective function is confined to decrease at every iteration of the algorithm.

Unfortunately, for cascading problems we cannot provide the tighter bound of Theorem 3 in Section 2.2. Even if we replace Assumption 1 with Assumption 2 in Proposition 7, there does not necessarily exists a *unique* set $I_d \in \mathcal{I}_d$ with $d = d_1 + d_2$ such that the map G_d^c, constructed as in Proposition 7, satisfies Assumption 2 (see also the construction of a set I_d that satisfies Assumption 1 in the proof of Proposition 7). However, one can relax Assumption 1 in Theorem 7 to Assumption 3 and replace the right-hand side of (28) according to Theorem 4. To ensure that the obtained solution violates the removed constraints, thus satisfying the second part of Assumption 3, we can follow the sampling and discarding procedure outlined in [8]. Removing a sample according to this procedure results in a reduction in the objective value of the optimization problem involved. In the cascading set-up, however, we have multiple objective functions and since both problems in the cascade are based on the same samples $\{\delta_i\}_{i=1}^m$, removing a sample affects the constraints in both problems. If for example we are interested, as in most applications, in the value of the last problem in the cascade, then removing a sample does not necessarily lead to a reduction in that objective value, since it may result in a different solution of the first problem in the cascade, which in turn affects the solution of the second problem. To incorporate this requirement in the removal procedure, we can eliminate a sample only if it results in a reduction in the objective value of the subproblem of interest.

5. Discussion

In Section 2 we showed that any algorithm that satisfies some consistency assumption (Assumption 1 or some of its strengthened or relaxed versions) is PAC-T learnable. In other words, consistency is a sufficient condition for learnability of a fixed, but possibly unknown, target concept $T \in \mathcal{D}$. The results of Section 2 can be easily extended to eansure learnability of an entire concept class $\mathcal{C} \subseteq \mathcal{D}$, thus implying that the underlying algorithm is PAC in the sense of [13]. However, following [13], [11], having a concept class with finite VC dimension (see [13] for a concise definition), which is a measure of the "richness" of this class, is a sufficient condition for PAC learnability. Therefore, the analysis of Section 2 complements the standard learning theoretic results based on VC theory, since an algorithm that generates a consistent hypothesis can be PAC even if the underlying concept class has infinite VC dimension.

Note that we consider here a fixed, but possibly unknown probability measure $\mathbb{P}$. However, if we are interested in learning a concept class uniformly with respect to any measure in some given class, then finite VC dimension is both a sufficient and necessary condition for PAC learnability. In this case, any concept class for which a consistent algorithm exists, would also have finite VC

dimension. The results of Section 2 can be then useful to provide tighter bounds without relying on the computation of the VC dimension, which might be a difficult task.

It should be also noted that Theorem 1 has a VC theoretic counterpart. For concept classes with finite VC dimension this is known as the probability of one-sided constrained failure [12], [13], [15], and for $m \geq 8/\epsilon$ it is of the form

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\quad \text{there exists } I_{d_{VC}}\in\mathcal{I}_{d_{VC}} \text{ such that}$$
$$H_{I_{d_{VC}}} \text{ is consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_{d_{VC}}})>\epsilon\Big\}\leq 2\sum_{i=0}^{d_{VC}}\binom{2m}{i}2^{-\frac{\epsilon m}{2}}, \quad (30)$$

where d_{VC} denotes the VC dimension. Despite the similarities between (30) and (3), the proofs of the corresponding statements are fundamentally different. However, it is shown in [1] that Assumption 1 is satisfied with $d = d_{VC}$ for a specific concept class with finite VC dimension, namely the so called maximum class. Connections between (30) and constraint violation properties of optimization problems can be found in [15].

Following the analysis of Section 2 for a generic algorithm, in Section 3 it was shown how the problem of providing guarantees regarding the probability of constraint satisfaction can be thought of as the problem of learning a specific target concept $T = \Delta$ for an algorithm that involves solving some optimization and generates a consistent hypothesis. Different examples were studied (the scenario approach, the probabilistically robust design, cascading optimization) to illustrate that these problems share certain similarities, thus justifying the reason that we obtain probabilistic performance guarantees of similar nature. In all cases, the probability that the measure of constraint violation exceeds a given threshold $\epsilon \in (0,1)$, is bounded by some function $q(m,\epsilon)$ such that $\lim_{m\to\infty} q(m,\epsilon) = 0$. The quantity $q(m,\epsilon)$ is the confidence with which we can provide constraint violation guarantees. In many applications it is of importance to compute explicit sample complexity bounds, i.e. determine the number of samples m for which $q(m,\epsilon) \leq \beta$, for some confidence level $\beta \in (0,1)$. The reader is referred to [1], [7], [9], for explicit bounds related to the involved $q(m,\epsilon)$ functions, and to [15], [21] for further refinements.

6. Concluding remarks

In this paper we considered a compression learning paradigm for algorithms that satisfy some consistency assumption. We first showed how one can strengthen or relax this assumption and analyzed the implications on the learnability properties. We then concentrated on scenario based optimization problems and showed that one can provide guarantees regarding the probability of constraint violation by treating them as learning problems. In this context, we also showed how novel probabilistic feasibility guarantees can be provided for cascading optimization problems. These novel results demonstrate how compressed learning can prove useful for scenario based multi-objective and sequential optimization problems.

Appendix A: Proofs of Section 2

Proof of Theorem 1. Consider any $\epsilon \in (0,1)$. The left-hand side of (3) can be expressed as follows:

$$\mathbb{P}^m\Big\{\bigcup_{I_d\in\mathcal{I}_d}\big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\big\}\Big\}.$$

From the subadditivity of $\mathbb{P}^m$ it then follows that

$$\mathbb{P}^m\Big\{ \bigcup_{I_d\in\mathcal{I}_d} \big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\big\}\Big\}$$
$$\leq \sum_{I_d\in\mathcal{I}_d} \mathbb{P}^m\Big\{\big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\big\}\Big\}. \tag{31}$$

Without loss of generality fix $I_d=\{1,\dots,d\}\in\mathcal{I}_d$ and consider any $(\delta_1,\dots,\delta_d)$ in the set $\bar{\Delta}^d=\{(\delta_1,\dots,\delta_d)\in\Delta^d:\ d_{\mathbb{P}}(T,H_{I_d}\big(T,\{\delta_i\}_{i=1}^d\big))>\epsilon\}$. Note that, under Assumption 1, H_{I_d} is consistent with $\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}$. If $\bar{\Delta}^d$ is empty, then, the contribution of I_d to the right-hand-side of (31) becomes zero, otherwise, we have that

$$\begin{aligned}\mathbb{P}\Big\{\delta\in\Delta:\ H_{I_d} \text{ is consistent with } \big(\delta,\mathbb{1}_T(\delta)\big) \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\Big\}&\\ =\mathbb{P}\Big\{\delta\in\Delta:\ H_{I_d} \text{ is consistent with } \big(\delta,\mathbb{1}_T(\delta)\big)\Big\}&\\ =1-d_{\mathbb{P}}(T,H_{I_d})\leq 1-\epsilon,&\end{aligned} \tag{32}$$

where the first step follows from the fact that $d_{\mathbb{P}}(T,H_{I_d}\big(T,\{\delta_i\}_{i=1}^d\big))$ does not depend on δ but only on $(\delta_1,\dots,\delta_d)\in\bar{\Delta}^d$, and the second step follows from the definition of a consistent hypothesis (Definition 2).

Since the samples are extracted independently we have that

$$\begin{aligned}\mathbb{P}^{m-d}\Big\{(\delta_{d+1},\dots,\delta_m)\in\Delta^{m-d}:\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=d+1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\Big\}&\\ =\prod_{j=d+1}^m \mathbb{P}\Big\{\delta_j\in\Delta:\ H_{I_d} \text{ is consistent with } \big(\delta_j,\mathbb{1}_T(\delta_j)\big) \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\Big\}&\\ \leq(1-\epsilon)^{m-d}.&\end{aligned} \tag{33}$$

Since (33) holds for any $(\delta_1,\dots,\delta_d)\in\bar{\Delta}^d$, we can rewrite the first quantity in (33) using the relevant conditional probability measure, denoted by Prob. We then have

$$\begin{aligned}\text{Prob}\Big\{\big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\big\}&\\ \Big|\big\{(\delta_1,\dots,\delta_d)\in\bar{\Delta}^d\big\}\Big\}\leq(1-\epsilon)^{m-d}.&\end{aligned} \tag{34}$$

Integrating with respect to the (conditional) probability of extracting a d-multisample $(\delta_1,\dots,\delta_d)$ from the set $\bar{\Delta}^d$ we get

$$\begin{aligned}\mathbb{P}^m\Big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\Big\}&\\ \leq\int_{\Delta^d}(1-\epsilon)^{m-d}\mathbb{1}_{\big\{(\delta_1,\dots,\delta_d)\in\bar{\Delta}^d\big\}}\mathbb{P}^d\big(\{\mathrm{d}\delta_i\}_{i\in I_d}\big)\leq(1-\epsilon)^{m-d},&\end{aligned} \tag{35}$$

for $I_d = \{1, \dots, d\} \in \mathcal{I}_d$. A similar reasoning can be applied to any $I_d \in \mathcal{I}_d$, which by equation (31) leads to:

$$\mathbb{P}^m\Big\{ \bigcup_{I_d \in \mathcal{I}_d} \big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\big\}\Big\}$$
$$\leq \sum_{I_d \in \mathcal{I}_d} \max\big\{0, (1-\epsilon)^{m-d}\big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}, \tag{36}$$

and concludes the proof. □

Proof of Theorem 2. Consider any $\epsilon \in (0,1)$. Under Assumption 1, let $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ be such that the hypothesis $H_{m_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in m_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. We then have that

$$\begin{aligned}
&\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_{m_d}) > \epsilon\Big\} \\
&= \mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_{m_d}) > \epsilon\Big\} \\
&= \mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{m_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{and } d_{\mathbb{P}}(T, H_{m_d}) > \epsilon\Big\},
\end{aligned} \tag{37}$$

where the last equality follows from Assumption 1. Now since the last term is upper bounded by

$$\begin{aligned}
\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ &\text{there exists } I_d \in \mathcal{I}_d \text{ such that} \\
&H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\},
\end{aligned} \tag{38}$$

by Theorem 1, we have that

$$\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_{m_d}) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}. \tag{39}$$

Set $q(m, \epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$. Since $\binom{m}{d} \leq \left(\frac{me}{d}\right)^d$ (Lemma 4.3 of [13]), we have that $\lim_{m\to\infty} q(m, \epsilon)$ $\leq \lim_{m\to\infty} \left(\frac{me}{d}\right)^d (1-\epsilon)^{m-d} = 0$. Therefore, $\lim_{m\to\infty} q(m, \epsilon) = 0$. Construct then algorithm $\big\{A_m\big\}_{m \geq d}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ takes as input a labeled m-multisample and returns a hypothesis $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ such that $H_m = H_{m_d}$. By Definition 4, algorithm $\big\{A_m\big\}_{m \geq d}$ is PAC-T. □

Proof of Proposition 1. We first show that $\mathbb{P}^m\big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ \Delta^m \setminus \cup_{I_d \in \mathcal{I}_d} S_{I_d}\big\} = 0$. It is equivalent to show that $\cup_{I_d \in \mathcal{I}_d} S_{I_d} = \Delta^m$ up to a set of measure zero. Clearly, $\cup_{I_d \in \mathcal{I}_d} S_{I_d} \subseteq \Delta^m$. Therefore, it suffices to show that $\cup_{I_d \in \mathcal{I}_d} S_{I_d} \supseteq \Delta^m$, i.e. if $(\delta_1, \dots, \delta_m) \in \Delta^m$ then there exists $I_d \in \mathcal{I}_d$ such that $(\delta_1, \dots, \delta_m) \in S_{I_d}$. With $\mathbb{P}^m$-probability one, the last statement follows from Assumption 2 and the definition of S_{I_d}.

It remains to show that $S_{I_d^1} \cap S_{I_d^2} = \emptyset$ for all $I_d^1, I_d^2 \in \mathcal{I}_d$ with $I_d^1 \neq I_d^2$. For the sake of contradiction assume that there exist $I_d^1, I_d^2 \in \mathcal{I}_d$ with $I_d^1 \neq I_d^2$ such that $S_{I_d^1} \cap S_{I_d^2} \neq \emptyset$. By the

definition of $S_{I_d^1}, S_{I_d^2}$, this implies that there exists $(\delta_1, \dots, \delta_m) \in \Delta^m$ such that both hypotheses $H_{I_d^1}$ and $H_{I_d^2}$ are consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. However, by Assumption 2 for any m-multisample $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ there exists a unique $I_d \in \mathcal{I}_d$ such that the corresponding hypothesis is consistent with respect to the m-multisample, establishing a contradiction. □

Proof of Proposition 2. Without loss of generality fix $I_d = \{1, \dots, d\} \in \mathcal{I}_d$ and consider any $(\delta_1, \dots, \delta_d) \in \Delta^d$. Denote by $\alpha\big(\{\delta_i\}_{i\in I_d}\big) = d_{\mathbb{P}}(T, H_{I_d})$ the error between the hypothesis H_{I_d} and the target concept T, and recall that, under the first part of Assumption 2, H_{I_d} is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}$. We have that

$$\begin{aligned}&\mathbb{P}\Big\{\delta \in \Delta :\ H_{I_d} \text{ is consistent with } \big(\delta, \mathbb{1}_T(\delta)\big)\Big\}\\ &\qquad = 1 - d_{\mathbb{P}}(T, H_{I_d}) = 1 - \alpha\big(\{\delta_i\}_{i\in I_d}\big).\end{aligned} \tag{40}$$

Since the samples are extracted independently we have that

$$\begin{aligned}&\mathbb{P}^{m-d}\Big\{(\delta_{d+1}, \dots, \delta_m) \in \Delta^{m-d} :\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=d+1}^m\Big\}\\ &\quad = \prod_{j=d+1}^m \mathbb{P}\Big\{\delta_j \in \Delta :\ H_{I_d} \text{ is consistent with } \big(\delta_j, \mathbb{1}_T(\delta_j)\big)\Big\}\\ &\quad = \Big(1 - \alpha\big(\{\delta_i\}_{i\in I_d}\big)\Big)^{m-d}.\end{aligned} \tag{41}$$

Integrating over $\big\{(\delta_1, \dots, \delta_d) \in \Delta^d\big\}$ we get

$$\begin{aligned}&\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m :\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\Big\}\\ &\qquad = \int_{\Delta^d} \Big(1 - \alpha\big(\{\delta_i\}_{i\in I_d}\big)\Big)^{m-d} \mathbb{P}^d\big(\{\mathrm{d}\delta_i\}_{i\in I_d}\big)\\ &\qquad = \int_0^1 (1-\alpha)^{m-d} F(\mathrm{d}\alpha),\end{aligned} \tag{42}$$

where the last equality is due to a change of variables and $F(\alpha)$ is defined by (5) and denotes the probability distribution of the error $d_{\mathbb{P}}(T, H_{I_d})$.

Since $S_{I_d} = \big\{(\delta_1, \dots, \delta_m) \in \Delta^m \mid H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big\}$, (42) implies that

$$\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in S_{I_d}\Big\} = \int_0^1 (1-\alpha)^{m-d} F(\mathrm{d}\alpha). \tag{43}$$

Under Assumption 2, Proposition 1 holds. Therefore, we have that S_{I_d}, $I_d \in \mathcal{I}_d$ form a partition of Δ^m up to a set of measure zero. Hence,

$$\sum_{I_d \in \mathcal{I}_d} \mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in S_{I_d}\Big\} = 1. \tag{44}$$

No set S_{I_d}, $I_d \in \mathcal{I}_d$ is more likely than the others, therefore, $\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in S_{I_d}\Big\}$ is the same for all $I_d \in \mathcal{I}_d$. This fact together with (44) implies that $\binom{m}{d}\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in S_{I_d}\Big\} = 1$.

The last statement together with (43) leads to

$$\binom{m}{d}\int_0^1 (1-\alpha)^{m-d}F(\mathrm{d}\alpha) = 1. \tag{45}$$

As shown in [8], there is a unique $F(\cdot)$ that satisfies (45). Integration by parts shows that $F(\alpha) = \alpha^d$ satisfies (45) and concludes the proof. □

Proof of Theorem 3. Consider any $\epsilon \in (0,1)$. Fix any $I_d \in \mathcal{I}_d$ and denote by $\alpha\big(\{\delta_i\}_{i\in I_d}\big) = d_{\mathbb{P}}(T, H_{I_d})$ the error between the hypothesis H_{I_d} and the target concept T. We then have that

$$\mathbb{P}^m\Big\{\bigcup_{I_d\in\mathcal{I}_d}\big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{I_d})>\epsilon\big\}\Big\} \tag{46}$$

$$= \binom{m}{d}\int_{\Delta^d}(1-\alpha\big(\{\delta_i\}_{i\in I_d}\big))^{m-d}\mathbb{1}_{\big\{\big(\{\delta_i\}_{i\in I_d}\big)\in\Delta^d:\ \alpha\left(\{\delta_i\}_{i\in I_d}\right)>\epsilon\big\}}\mathbb{P}^d\big(\{\mathrm{d}\delta_i\}_{i\in I_d}\big) \tag{47}$$

$$= \binom{m}{d}\int_\epsilon^1 (1-\alpha)^{m-d}F(\mathrm{d}\alpha) \tag{48}$$

$$= \binom{m}{d}\int_\epsilon^1 (1-\alpha)^{m-d}d\alpha^{d-1}\mathrm{d}\alpha \tag{49}$$

$$= \sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}, \tag{50}$$

where (47) follows from (46) using (41), (48) follows from (47) by a change of variables and (49) follows from the fact that, under Assumption 2, Proposition 2 implies that $F(\alpha) = \alpha^d$. Equality (50) follows by repeated integration by parts (see also p. 1219 of [8]).

Under Assumption 2, let $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ be the unique set of indices such that the hypothesis $H_{m_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in m_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Since m_d is unique, (46) is equal to

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ H_{m_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T,H_{m_d})>\epsilon\Big\}, \tag{51}$$

which based on (37) (see proof of Theorem 2) is equal to $\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ d_{\mathbb{P}}(T,H_{m_d})>\epsilon\Big\}$. Therefore,

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ d_{\mathbb{P}}(T,H_{m_d})>\epsilon\Big\} = \sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}. \tag{52}$$

Set $q(m,\epsilon) = \sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}$. As shown in Lemma 4.3 of [13], $\binom{m}{i} \le \left(\frac{me}{i}\right)^i$ for any $i\in\mathbb{Z}$. Therefore, $q(m,\epsilon) = \sum_{i=0}^{d-1}\epsilon^i\binom{m}{i}(1-\epsilon)^{m-i} \le \sum_{i=0}^{d-1}\epsilon^i\left(\frac{me}{i}\right)^i(1-\epsilon)^{m-i}$. Following the proof of Theorem 2, every term in last summation is such that $\lim_{m\to\infty}\left(\frac{me}{i}\right)^i(1-\epsilon)^{m-i} = 0$. Therefore,

$\lim_{m\to\infty} q(m,\epsilon) = 0$. Construct then algorithm $\{A_m\}_{m\geq d}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ takes as input a labeled m-multisample and returns a hypothesis $H_m = A_m\big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i=1}^m\big)$ such that $H_m = H_{m_d}$. By Definition 4, algorithm $\{A_m\}_{m\geq d}$ is PAC-T. □

Proof of Theorem 4. Fix any $r \in \mathbb{N}$ and $I_r \in \mathcal{I}_r$, and under the second part of Assumption 3, let $m_d^r\Big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in\{1,\ldots,m\}\setminus I_r}\Big)$ be a set of d indices such that $H_{m_d^r} = G_d\big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in m_d^r}\big)$ is consistent with $\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in\{1,\ldots,m\}\setminus I_r}$. Denote then by $\alpha\big(\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r}\big) = d_{\mathbb{P}}(T, H_{m_d^r})$ the error between the hypothesis $H_{m_d^r}$ and the target concept T. We then have that

$$\begin{aligned}&\mathbb{P}\Big\{\delta \in \Delta :\ H_{m_d^r} \text{ is not consistent with } \big(\delta, \mathbb{1}_T(\delta)\big)\Big\}\\ &\qquad = d_{\mathbb{P}}(T, H_{m_d^r}) = \alpha\big(\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r}\big).\end{aligned} \tag{53}$$

Since the samples are extracted independently we have that

$$\begin{aligned}&\mathbb{P}^r\Big\{\{\delta_j\}_{j\in I_r} \in \Delta^r :\ H_{m_d^r} \text{ is not consistent with } \{(\delta_j, \mathbb{1}_T(\delta_j))\}_{j\in I_r}\Big\}\\ &\quad = \prod_{j\in I_r} \mathbb{P}\Big\{\delta_j \in \Delta :\ H_{m_d^r} \text{ is not consistent with } \big(\delta_j, \mathbb{1}_T(\delta_j)\big)\Big\}\\ &\quad = \alpha\big(\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r}\big)^r.\end{aligned} \tag{54}$$

Consider any $\epsilon \in (0,1)$. We then have that

$$\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ H_{m_d^r} \text{ is not consistent with } \{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in I_r} \text{ and } d_{\mathbb{P}}(T, H_{m_d^r}) > \epsilon\Big\} \tag{55}$$

$$= \int_{\Delta^{m-r}} \alpha\big(\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r}\big)^r \mathbb{1}_{\left\{\alpha\left(\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r}\right) > \epsilon\right\}} \mathbb{P}^{m-r}\big(\{\mathrm{d}\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r}\big) \tag{56}$$

$$= \int_{\epsilon}^{1} \alpha^r \bar{F}(\mathrm{d}\alpha), \tag{57}$$

where (56) follows from (55) using (54), (57) follows from (56) by a change of variables and $\bar{F}(\cdot)$ is the probability distribution of the error $d_{\mathbb{P}}(T, H_{m_d^r})$, i.e.

$$\bar{F}(\alpha) = \mathbb{P}^{m-r}\big\{\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r} \in \Delta^{m-r} :\ d_{\mathbb{P}}(T, H_{m_d^r}) \leq \alpha\big\}. \tag{58}$$

Notice the difference between (58) and (5); the latter is the distribution of the error between the target concept and the hypothesis generated using all elements of the multisample, whereas the former is the distribution of the error between the target concept and the hypothesis using only d out of the $m-r$ elements of the multisample.

Construct the algorithm $\{A_{m-r}\}_{m-r\geq d}$, where $A_{m-r} : [\Delta \times \{0,1\}]^{m-r} \to \mathcal{D}$ takes as input a labeled $m-r$-multisample and returns a hypothesis $H_{m-r} = A_{m-r}\big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in\{1,\ldots,m\}\setminus I_r}\big)$ such that $H_{m-r} = H_{m_d^r}$. By Theorem 2 with $m-r$ in place of m and α in place of ϵ, we have that the constructed algorithm is PAC-T, hence

$$\mathbb{P}^{m-r}\Big\{\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r} \in \Delta^{m-r} :\ d_{\mathbb{P}}(T, H_{m_d^r}) > \alpha\Big\} \leq \binom{m-r}{d}(1-\alpha)^{m-r-d}. \tag{59}$$

By (58), (59), we then have that $\bar{F}(\alpha) \geq 1 - \binom{m-r}{d}(1-\alpha)^{m-r-d}$ for any $\alpha \in [0,1]$. The last statement together with (55)-(57) leads to

$$\begin{aligned}
&\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ H_{m_d^r} \text{ is not consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r} \text{ and } d_{\mathbb{P}}(T, \bar{H}_{m-r}) > \epsilon\Big\} \\
&\quad = \int_\epsilon^1 \alpha^r \bar{F}(\mathrm{d}\alpha) && (60)\\
&\quad \leq \int_\epsilon^1 (m-r-d)\binom{m-r}{d}\alpha^r\big(1-\alpha\big)^{m-r-d-1}\mathrm{d}\alpha && (61)\\
&\quad = \binom{m-r}{d}\frac{1}{\binom{m-d}{r}}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i}, && (62)
\end{aligned}$$

where the inequality in (61) follows from (60) due to the fact that $\bar{F}(\alpha) \geq 1 - \binom{m-r}{d}(1-\alpha)^{m-r-d}$ and is based on standard integral arguments (see also the proof of Theorem 2.1 in [18]). Moreover, (62) follows from (61) by repeated integration by parts.

Denote now by $\bar{I}_r \in \mathcal{I}_r$ the set of indices for which the third part of Assumption 3 is satisfied. Let then $\bar{m}_d^r\Big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\ldots,m\}\setminus\bar{I}_r}\Big)$ be a set of d indices such that the hypothesis $H_{\bar{m}_d^r} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\bar{m}_d^r}\big)$ is not consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\bar{I}_r}$. We thus have

$$\begin{aligned}
&\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ d_{\mathbb{P}}(T, H_{\bar{m}_d^r}) > \epsilon\Big\} \\
&\quad \leq \mathbb{P}^m\Big\{ \bigcup_{I_r\in\mathcal{I}_r} \big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ H_{m_d^r} \text{ is not consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r} \\
&\qquad\qquad \text{and } d_{\mathbb{P}}(T, H_{m_d^r}) > \epsilon\big\}\Big\} && (63)\\
&\quad \leq \sum_{I_r\in\mathcal{I}_r} \mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ H_{m_d^r} \text{ is not consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r} \\
&\qquad\qquad \text{and } d_{\mathbb{P}}(T, H_{m_d^r}) > \epsilon\Big\}, && (64)\\
&\quad = \binom{m}{r}\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ H_{m_d^r} \text{ is not consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r} \\
&\qquad\qquad \text{and } d_{\mathbb{P}}(T, H_{m_d^r}) > \epsilon\Big\}, && (65)
\end{aligned}$$

where (64) is due to the subadditivity of $\mathbb{P}^m$. By (62), (65) we have that

$$\begin{aligned}
&\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m) \in \Delta^m :\ d_{\mathbb{P}}(T, H_{\bar{m}_d^r}) > \epsilon\Big\} \\
&\quad \leq \binom{m}{r}\binom{m-r}{d}\frac{1}{\binom{m-d}{r}}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i} && (66)\\
&\quad = \binom{m}{d}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i}. && (67)
\end{aligned}$$

Set $q(m,\epsilon) = \binom{m}{d}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i}$. Similarly to the last part of the proof of Theorem 3, $\lim_{m\to\infty} q(m,\epsilon) = 0$. Construct then algorithm $\big\{A_m\big\}_{m\geq d+r}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$

takes as input a labeled m-multisample and returns a hypothesis $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ such that $H_m = H_{\bar{m}_d^r}$. By Definition 4, algorithm $\big\{A_m\big\}_{m\geq d+r}$ is PAC-T. $\square$

Proof of Theorem 5. The proof of Theorem 5 follows the same lines with the proof of Theorem 4 up to equation (59). Instead of (59), Theorem 3 with $m-r$ in place of m and α in place of ϵ implies that

$$\mathbb{P}^{m-r}\Big\{\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_r} \in \Delta^{m-r} :\ d_{\mathbb{P}}(T, H_{m_d^r}) > \alpha\Big\} = \sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}. \tag{68}$$

By (58), (68), we then have that $\bar{F}(\alpha) = 1 - \sum_{i=0}^{d-1}\binom{m}{i}\alpha^i(1-\alpha)^{m-i}$ for any $\alpha\in[0,1]$. The last statement together with (55)-(57) leads to

$$\begin{aligned}
&\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m :\ H_{m_d^r} \text{ is not consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r} \text{ and } d_{\mathbb{P}}(T, H_{m_d^r}) > \epsilon\Big\}\\
&\quad= \int_\epsilon^1 \alpha^r \bar{F}(\mathrm{d}\alpha) &(69)\\
&\quad= \int_\epsilon^1 d\binom{m-r}{d}\alpha^r\alpha^{r-1}\big(1-\alpha\big)^{m-r-d}\mathrm{d}\alpha &(70)\\
&\quad= \frac{d\binom{m-r}{d}}{(r+d)\binom{m}{r+d}}\sum_{i=0}^{r+d-1}\binom{m}{i}\alpha^i(1-\alpha)^{m-i}, &(71)
\end{aligned}$$

where the equality in (70) (compare with the inequality in (61)) follows from (69) due to the fact that $\bar{F}(\alpha) = 1 - \sum_{i=0}^{d-1}\binom{m}{i}\alpha^i(1-\alpha)^{m-i}$ and is based on standard integral arguments (see also the proof of Theorem 2.1 in [18]). Moreover, (71) follows from (70) by repeated integration by parts.

Construct an algorithm as shown above (63) and follow the same arguments with (63)-(65). By (71), (65) we have that

$$\begin{aligned}
&\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m :\ d_{\mathbb{P}}(T, H_{\bar{m}_d^r}) > \epsilon\Big\}\\
&\quad\leq \binom{m}{r}\frac{d\binom{m-r}{d}}{(r+d)\binom{m}{r+d}}\sum_{i=0}^{r+d-1}\binom{m}{i}\alpha^i(1-\alpha)^{m-i} &(72)\\
&\quad= \binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\alpha^i(1-\alpha)^{m-i}. &(73)
\end{aligned}$$

Set $q(m,\epsilon) = \binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\alpha^i(1-\alpha)^{m-i}$. Similarly to the last part of the proof of Theorem 3, $\lim_{m\to\infty} q(m,\epsilon) = 0$. Construct then algorithm $\big\{A_m\big\}_{m\geq d+r}$, where $A_m : [\Delta\times\{0,1\}]^m \to \mathcal{D}$ takes as input a labeled m-multisample and returns a hypothesis $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ such that $H_m = H_{\bar{m}_d^r}$. By Definition 4, algorithm $\big\{A_m\big\}_{m\geq d+r}$ is PAC-T. $\square$

Appendix B: Proofs of Sections 3

Proof of Theorem 6. Under Assumption 1, the hypothesis $H_{m_d} = \big\{\delta\in\Delta :\ g(x_d(\{\delta_i\}_{i\in m_d}), \delta)\leq 0\big\}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. This implies that $x_d(\{\delta_i\}_{i\in m_d})$ belongs to the feasibility

region of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$. Consider an algorithm $\{A_m\}_{m\geq d}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ is such that $H_m = A_m\big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i=1}^m\big)$ with $H_m = \{\delta \in \Delta :\ g(x_m(\{\delta_i\}_{i=1}^m), \delta) \leq 0\}$. Moreover, by the theorem hypothesis we have that $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$, which entails that $H_m = H_{m_d} = G_d\big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in m_d}\big)$, for G_d defined according to (11). Theorem 2 implies then that $\{A_m\}_{m\geq d}$ is PAC-T with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$. The latter, together with the fact that, since $T = \Delta$, $d_{\mathbb{P}}(T, H_m) = \mathbb{P}\big(\{\delta \in \Delta :\ g(x_m(\{\delta_i\}_{i=1}^m), \delta) > 0\}\big)$, leads to (13). □

Proof of Proposition 3. Fix $d = \zeta$ and consider $m \geq d$. By the definition of the support constraints, and under Assumption 5, with $\mathbb{P}^m$-probability one, there exists $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ such that $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$ [8], where x_m, x_d denote the unique (under Assumption 5) minimizers of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ and $\mathcal{P}_1[\{\delta_i\}_{i\in m_d}]$, respectively. The solution $x_d(\{\delta_i\}_{i\in m_d})$ satisfies all constraints that correspond to samples whose indices are not included in m_d, otherwise we would not have $x_d(\{\delta_i\}_{i\in m_d}) = x_m$. In other words, $g(x_d(\{\delta_i\}_{i\in m_d}), \delta_i) \leq 0$ for all $i \in \{1,\ldots,m\} \setminus m_d$. But, since $x_d(\{\delta_i\}_{i\in m_d})$ is the optimal solution of $\mathcal{P}_1[\{\delta_i\}_{i\in m_d}]$ it will satisfy its constraints, i.e. $g(x_d(\{\delta_i\}_{i\in m_d}), \delta_i) \leq 0$ for all $i \in m_d$. Therefore, $g(x_d(\{\delta_i\}_{i\in m_d}), \delta_i) \leq 0$ for all $i \in \{1,\ldots,m\}$ and since $H_{m_d} = \{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in m_d}), \delta) \leq 0\}$, we have that $\mathbb{1}_{H_{m_d}}(\delta_i) = 1$, for all $i = 1,\ldots,m$. The last statement together with the fact that $T = \Delta$ implies that the hypothesis $H_{m_d} = G_d\big(\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in m_d}\big)$ is consistent with $\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i=1}^m$, thus showing that the second part of Assumption 1 is satisfied.

It remains to show the first part of Assumption 1. For any $I_d \in \mathcal{I}_d$, since $x_d(\{\delta_i\}_{i\in I_d})$ is the minimizer of optimal solution of $\mathcal{P}_1[\{\delta_i\}_{i\in I_d}]$ it will satisfy its constraints, i.e. $g(x_d(\{\delta_i\}_{i\in m_d}), \delta_i) \leq 0$ for all $i \in I_d$. By definition, it then follows that H_{I_d} is consistent with $\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i\in I_d}$. □

Proof of Corollary 1. Under Assumption 5, Proposition 3 shows that $d = \zeta$, G_d satisfy Assumption 1. Let then $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ be a set of indices for which the consistency requirement of Assumption 1 is satisfied. Moreover, as shown in the proof of Proposition 3, $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$. Theorem 6 leads then to (15) and concludes the proof. □

Proof of Proposition 4. Assume that $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ has exactly $d = \zeta \leq m$ support constraints with $\mathbb{P}^m$-probability one. Under Assumption 5, and since the number of support constraints is bounded, Proposition 3 implies that Assumption 1 is satisfied. However, for the sake of contradiction assume that the second part of Assumption 2 is not satisfied. Therefore, there exist $I_d^1, I_d^2 \in \mathcal{I}_d$ with $I_d^1 \neq I_d^2$ such that the hypotheses $H_{I_d^1}$ and $H_{I_d^2}$ are both consistent with $\{(\delta_i, \mathbb{1}_T(\delta_i))\}_{i=1}^m$. $H_{I_d^1}$ and $H_{I_d^2}$ are constructed according to (11) based on the minimizers $x_d(\{\delta_i\}_{i\in I_d^1})$ and $x_d(\{\delta_i\}_{i\in I_d^2})$ of problems $\mathcal{P}_1[\{\delta_i\}_{i\in I_d^1}]$ and $\mathcal{P}_1[\{\delta_i\}_{i\in I_d^2}]$, respectively. By the definition of consistency and the form of $H_{I_d^1}$, $H_{I_d^2}$, we have that $x_d(\{\delta_i\}_{i\in I_d^1})$ and $x_d(\{\delta_i\}_{i\in I_d^2})$ satisfy also all constraints corresponding to $\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_d^1}$ and $\{\delta_i\}_{i\in\{1,\ldots,m\}\setminus I_d^1}$, respectively. Therefore, and under the uniqueness part of Assumption 5, $x_d(\{\delta_i\}_{i\in I_d^1}) = x_d(\{\delta_i\}_{i\in I_d^2}) = x_m(\{\delta_i\}_{i=1}^m)$, where x_m is the minimizer of $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$.

By the definition of the support constraints (see Definition 4 in [7] and discussion in Section 3.2) and under the assumption that $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ has exactly d support constraints, the last statement implies that the constraints that correspond to samples with indices in I_d^1 and I_d^2 are support constraints. Moreover, the fact that $I_d^1 \neq I_d^2$ would imply that there exist at least one index that does not belong to both I_d^1 and I_d^2. Since $|I_d^1| = |I_d^2| = d$, the last statement implies that the number of support constraints would be greater than or equal to $d+1$, thus contradicting the assumption that we have exactly d support constraints, proving the second part of Assumption 2. To conclude

the proof it remains to show the first part of Assumption 2; this is the same with the first statement of Assumption 1 and can be shown as in the proof of Proposition 3. □

Proof of Corollary 2. Under Assumption 5, and since $\mathcal{P}_1[\{\delta_i\}_{i=1}^m]$ has $d = \zeta \leq m$ support constraints with $\mathbb{P}^m$-probability one, Proposition 4 shows that d, G_d satisfy Assumption 2. Similarly to the proof of Corollary 1, let $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ be the unique set of indices for which the consistency requirement of Assumption 2 is satisfied. Moreover, as shown in the proof of Proposition 3, $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$. Theorem 6 leads then to (16) and concludes the proof. □

Proof of Proposition 5. We first show that with $\mathbb{P}^m$-probability one, there exists a unique set $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ with $d = 2n_\delta$ such that $B(p_d(\{\delta_i\}_{i\in m_d})) = B(p_m(\{\delta\}_{i=1}^m))$. $B(p_d(\{\delta_i\}_{i\in m_d}))$ is the minimum volume hyper-rectangle that contains $d = 2n_\delta$ samples of the uncertainty with indices given by m_d. Clearly, $B(p_d(\{\delta_i\}_{i\in I_d})) \subseteq B(p_m)$ for any $I_d \in \mathcal{I}_d$. Therefore, it suffices to show that there exists $I_d \in \mathcal{I}_d$ such that $B(p_m) = B(p_d(\{\delta_i\}_{i\in m_d}))$. Let $B(p_m) = \times_{\ell=1}^{n_\delta}[\underline{p}_m^\ell,\ \overline{p}_m^\ell]$, where $\underline{p}_m^\ell$, $\overline{p}_m^\ell$ denote the ℓ-th elements of $\underline{p}_m$ and $\overline{p}_m$, respectively. By inspection of $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m]$, for all $\ell = 1, \ldots, n_\delta$, $\underline{p}_m^\ell = \min_{i=1,\ldots,m} \delta_i^\ell$ and $\overline{p}_m^{*,\ell} = \max_{i=1,\ldots,m} \delta_i^\ell$, where δ_i^ℓ denotes the ℓ-th element of sample i.

With $\mathbb{P}^m$-probability one $\widetilde{\mathcal{P}}_2[\{\delta_i\}_{i=1}^m]$ admits a unique solution. Let $\underline{i}^\ell = \arg\min_{i=1,\ldots,m} \delta_i^\ell$ and $\overline{i}^\ell = \arg\max_{i=1,\ldots,m} \delta_i^\ell$, for $\ell = 1, \ldots, n_\delta$. Consider then the set of indices $m_d = \big\{\{\underline{i}^\ell, \overline{i}^\ell\}_{\ell=1}^{n_\delta}\big\}$. With $\mathbb{P}^m$-probability one, m_d is unique, $|m_d| = 2n_\delta$ and by construction (see the definition of $B(p_m)$) $B(p_d(\{\delta_i\}_{i\in m_d})) = B(p_m)$.

Fix $d = 2n_\delta$ and consider $m \geq d$. Let $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ be the unique set of indices for which $B(p_d(\{\delta_i\}_{i\in m_d})) = B(p_m)$ and denote by $\mathcal{X}_{m_d} = \{x \in \mathcal{X} :\ g(x,\delta) \leq 0,\ \forall \delta \in B(p_d(\{\delta_i\}_{i\in m_d}))\}$ the feasibility region of $\mathcal{P}_2[\{\delta_i\}_{i\in m_d}]$, which is non-empty by Assumption 6. Let then $x_d(\{\delta_i\}_{i\in m_d})$ be a minimizer of $\mathcal{P}_2[\{\delta_i\}_{i\in m_d}]$. By construction $x_d(\{\delta_i\}_{i\in m_d}) \in \mathcal{X}_{m_d}$, so it would satisfy all constraints of $\mathcal{P}_2[\{\delta_i\}_{i\in m_d}]$. Therefore, $g(x_d(\{\delta_i\}_{i\in m_d}), \delta) \leq 0$ for all $\delta \in B(p_d(\{\delta_i\}_{i\in m_d}))$. Since $B(p_d(\{\delta_i\}_{i\in m_d})) = B(p_m)$, the last statement is equivalent to

$$g(x_d(\{\delta_i\}_{i\in m_d}), \delta) \leq 0 \text{ for all } \delta \in B(p_m). \tag{74}$$

The hypothesis H_{m_d} is given by $H_{m_d} = \big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in m_d}), \delta) \leq 0\big\}$. By (74), this implies that

$$\mathbb{1}_{H_{m_d}}(\delta) = 1, \text{ for all } \delta \in B(p_m). \tag{75}$$

Since $B(p_m)$ contains all samples $\delta_1, \ldots, \delta_m$, the last statement implies that $\mathbb{1}_{H_{m_d}}(\delta_i) = 1$ for all $i = 1, \ldots, m$. The last statement together with the fact that $T = \Delta$ implies that the hypothesis $H_{m_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in m_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$, thus showing that the second part of Assumption 1 is satisfied. To conclude the proof it remains to show the first part of Assumption 1; this can be done as in the proof of Proposition 3. □

Proof of Corollary 3. Under Assumption 6, Proposition 5 shows that $d = 2n_\delta$, G_d satisfy Assumption 1. Let then $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ be the unique (under Proposition 5) set of indices for which the consistency requirement of Assumption 1 is satisfied. Moreover, as shown in the proof of Proposition 3, $B(p_m(\{\delta_i\}_{i=1}^m)) = B(p_d(\{\delta_i\}_{i\in m_d}))$, which implies that $\mathcal{X}_{m_d} = \mathcal{X}_m$. Due to the uniqueness part of Assumption 6 we then have that $x_d(\{\delta_i\}_{i\in m_d}) = x_m(\{\delta_i\}_{i=1}^m)$. Theorem 6 leads then to (19) and concludes the proof. □

Proof of Proposition 6. If (22) is satisfied, then, with $\mathbb{P}^m$-probability one, for all $I_d \in \mathcal{I}_d$

$$\big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in I_d}), \delta) > 0\big\} = \big\{\delta \in \Delta :\ \delta \notin B(p_d(\{\delta_i\}_{i\in I_d}))\big\}. \tag{76}$$

As shown in the proof of Proposition 5, there exists a unique $m_d \in \mathcal{I}_d$ such that $B(p_d(\{\delta_i\}_{i\in m_d})) = B(p_m(\{\delta_i\}_{i=1}^m))$. For any $I_d \in \mathcal{I}_d$ with $I_d \neq m_d$ we have that $B(p_d(\{\delta_i\}_{i\in I_d})) \subset B(p_m(\{\delta_i\}_{i=1}^m)) = B(p_d(\{\delta_i\}_{i\in m_d}))$.

For the sake of contradiction assume that second part of Assumption 2 is not satisfied. This implies that there exists $I_d \in \mathcal{I}_d$ with $I_d \neq m_d$ such that H_{I_d} is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Since $H_{I_d} = \{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in I_d}), \delta) \leq 0\}$ (x_d is a minimizer of $\mathcal{P}_2[\{\delta_i\}_{i\in I_d}]$), consistency implies that $g(x_d(\{\delta_i\}_{i\in I_d})), \delta_i) \leq 0$ for all $i = 1, \ldots, m$. Since $B(p_d(\{\delta_i\}_{i\in I_d})) \subset B(p_m)$, the last statement implies that there exists $\ell \in \{1, \ldots, m\} \setminus I_d$ such that $\delta_\ell \notin B(p_d(\{\delta_i\}_{i\in I_d}))$ and $g(x_d(\{\delta_i\}_{i\in I_d}), \delta_\ell) \leq 0$, i.e. there exists at least one uncertainty realization δ that is not contained in $B(p_d(\{\delta_i\}_{i\in I_d}))$ and does not lead to constraint violation. Therefore,

$$\big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in I_d}), \delta) > 0\big\} \subset \big\{\delta \in \Delta :\ \delta \notin B(p_d(\{\delta_i\}_{i\in I_d}))\big\}. \tag{77}$$

Equations (76) and (77) establish a contradiction, proving the second part of Assumption 2. To conclude the proof it remains to show the first part of Assumption 2; this is the same with the first statement of Assumption 1 and can be shown as in the proof of Proposition 3. □

Proof of Corollary 4. Under Assumption 6, and since (22) is satisfied with $\mathbb{P}^m$-probability one, by Proposition 6 we have that Assumption 2 is satisfied. Equation (23) results then from (6) in Theorem 3, following similar arguments with the proof of Corollary 3. □

Appendix C: Proofs of Sections 4

Proof of Proposition 7. Under Assumption 5, Assumption 1 is satisfied for $d_1 \in \mathbb{N}$, $G_{d_1} : [\Delta \times \{0,1\}]^{d_1} \to \mathcal{D}$. Then, there exists $m_{d_1}(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_{d_1}$ such that $H_{m_{d_1}} = G_{d_1}\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in m_{d_1}}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Since $H_{m_{d_1}} = \big\{\delta \in \Delta :\ g\big(x_{d_1}(\{\delta_i\}_{i\in m_{d_1}}), \delta\big) \leq 0\big\}$,

$$g\big(x_{d_1}(\{\delta_i\}_{i\in m_{d_1}}), \delta_i\big) \leq 0, \text{ for all } i = 1, \ldots, m. \tag{78}$$

Moreover, under Assumption 5, for all $x \in \mathcal{X}$, Assumption 1 is satisfied for $d_2 \in \mathbb{N}$, $\widetilde{G}_{d_2}[x] : [\Delta \times \{0,1\}]^{d_2} \to \mathcal{D}$. This implies that, for all $x \in \mathcal{X}$, there exists $m_{d_2}[x](\{\delta_i\}_{i=1}^m) \in \mathcal{I}_{d_2}$ such that the hypothesis $\widetilde{H}_{m_{d_2}[x]}[x] = \widetilde{G}_{d_2}[x]\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in m_{d_2}[x]}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Since $\widetilde{H}_{m_{d_2}[x]}[x] = \big\{\delta \in \Delta :\ \widetilde{g}\big(y_{d_2}[x](\{\delta_i\}_{i\in m_{d_2}[x]}, x, \delta)\big) \leq 0\big\}$, for any $x \in \mathcal{X}$,

$$\widetilde{g}\big(y_{d_2}[x](\{\delta_i\}_{i\in m_{d_2}[x]}, x, \delta_i)\big) \leq 0, \text{ for all } i = 1, \ldots, m. \tag{79}$$

Set $d = d_1 + d_2$ and consider $m \geq d$. Choose $m_d(\{\delta_i\}_{i=1}^m) \in \mathcal{I}_d$ such that $m_d(\{\delta_i\}_{i=1}^m) \supseteq m_{d_1}(\{\delta_i\}_{i=1}^m) \cup m_{d_2}[x_{d_1}(\{\delta_i\}_{i\in m_{d_1}(\{\delta_i\}_{i=1}^m)})](\{\delta_i\}_{i=1}^m)$ (we do not have equality since some indices may belong to both m_{d_1} and $m_{d_2}[x]$, implying that some constraints are of support for both problems in the cascade), where $x_{d_1}(\{\delta_i\}_{i\in m_{d_1}})$ is the minimizer of $\mathcal{P}[\{\delta_i\}_{i\in m_{d_1}}]$ that is used to construct $H_{m_{d_1}}$. For simplicity, as in (78), (79), we do not show the argument $(\{\delta_i\}_{i=1}^m)$ of m_{d_1}, $m_{d_2}[x_{d_1}(\{\delta_i\}_{i\in m_{d_1}})]$. As shown in the proof of Proposition 3, since $m_d \supseteq m_{d_1}$, $x_d(\{\delta_i\}_{i\in m_d}) = x_{d_1}(\{\delta_i\}_{i\in m_{d_1}})$. Therefore, (78) implies that $g\big(x_d(\{\delta_i\}_{i\in m_d}), \delta_i\big) \leq 0$, for all $i = 1, \ldots, m$. We also

have that $m_d \supseteq m_{d_2}[x_{d_1}(\{\delta_i\}_{i\in m_{d_1}})] = m_{d_2}[x_d(\{\delta_i\}_{i\in m_d})]$, where the last equality follows from the fact that $x_d(\{\delta_i\}_{i\in m_d}) = x_{d_1}(\{\delta_i\}_{i\in m_{d_1}})$. Similarly to the previous case, as shown in the proof of Proposition 3 we have that $y_d[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_d}) = y_{d_2}[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_{d_2}[x_d(\{\delta_i\}_{i\in m_d})]})$. By (79), $\widetilde{g}\big(y_d[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_d}), x_d(\{\delta_i\}_{i\in m_d}), \delta_i)\big) \leq 0$, for all $i = 1, \ldots, m$. Therefore, we have that

$$g\big(x_d(\{\delta_i\}_{i\in m_d}),\delta_i\big) \leq 0 \text{ and}$$
$$\widetilde{g}\big(y_d[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_d}), x_d(\{\delta_i\}_{i\in m_d}), \delta_i)\big) \leq 0, \text{ for all } i = 1, \ldots, m. \quad (80)$$

Since $T = \Delta$, (80), (27) imply that $G_d^c\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in m_d}\big) = H_{m_d} \cap \widetilde{H}_{m_d}[x_d(\{\delta_i\}_{i\in m_d})]$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. To conclude the proof it remains to show the first part of Assumption 1; this can be done as in the proof of Proposition 3. □

Proof of Theorem 7. Under Assumption 5, Proposition 7 implies that G_d^c satisfies Assumption 1. Then, there exists $m_d \in \mathcal{I}_d$ such that the hypothesis $H_{m_d}^c = \big\{\delta \in \Delta : \big(g(x_{m_d}, \delta) \leq 0\big) \text{ and } \big(\widetilde{g}(y_{m_d}[x_{m_d}], x_{m_d}, \delta) \leq 0\big)\big\}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Consider an algorithm $\big\{A_m\big\}_{m\geq d}$, where $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ is such that $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ with $H_m = \big\{\delta \in \Delta : \big(g(x_m, \delta) \leq 0\big) \text{ and } \big(\widetilde{g}(y_m[x_m], x_m, \delta) \leq 0\big)\big\}$. Under Assumption 5, following the proof of Proposition 3 we have that $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$ and $y_m[x_m](\{\delta_i\}_{i=1}^m) = y_d[x_d](\{\delta_i\}_{i\in m_d})$ and hence $H_m = H_{m_d}^c = G_d^c\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in m_d}\big)$. Theorem 2 implies then that $\big\{A_m\big\}_{m\geq d}$ is PAC-T with $q(m,\epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$. The latter, together with the fact that, since $T = \Delta$, $d_{\mathbb{P}}(T, H_m) = \mathbb{P}\Big(\delta \in \Delta : \big(g(x_m, \delta) > 0\big) \text{ or } \big(\widetilde{g}(y_m[x_m], x_m, \delta) > 0\big)\Big)$, leads to (28). □

References

[1] S. Floyd and M. Warmuth, "Sample compression, learnability, and the Vapnik-Chervonenkis dimension," *Machine Learning*, pp. 1–36, 1995.

[2] A. Ben-Tal, L. El-Ghaoui, and A. Nemirovski, *Robust Optimization*. Princeton Series in Applied Mathematics, 2009.

[3] A. Prekopa, *Stochastic Programming*. Cluwer Academic Publishers, Dordrecht, Boston, 1995.

[4] A. Shapiro, "Stochastic programming approach to optimization under uncertainty ," *Mathematical Programming, Series B*, vol. 112, pp. 183 – 183, 2008.

[5] A. Nemirovski and A. Shapiro, "Convex Approximations of Chance Constrained Programs," *Siam Journal on Control and Optimization*, vol. 17, no. 4, pp. 969 – 996, 2006.

[6] D. Bertsimas and M. Sim, "Tractable Approximations to Robust Conic Optimization Problems," *Mathematical Programming, Series B*, vol. 107, pp. 5–36, 2006.

[7] G. Calafiore and M. Campi, "The scenario approach to robust control design," *IEEE Transactions on Automatic Control*, vol. 51, no. 5, pp. 742–753, 2006.

[8] M. Campi and S. Garatti, "The exact feasibility of randomized solutions of uncertain convex programs," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1211–1230, 2008.

[9] G. Calafiore, "Random Convex Programs," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3427–3464, 2010.

[10] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory Probab. Appl.*, vol. 16, no. 2, pp. 264 – 280, 1971.

[11] V. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

[12] M. Anthony and N. Biggs, *Computational Learning Theory*. Cambridge Tracts in Theoretical Computer Science, 1992.

[13] M. Vidyasagar, *A Theory of Learning and Generalization*. London, U.K.: Springer-Verlag, 1997.

[14] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized Algorithms for Analysis and Control of Uncertain Systems*. Springer-Verlag, London, 2005.

[15] T. Alamo, R. Tempo, and E. Camacho, "Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2545 – 2559, 2009.

[16] K. Margellos, P. Goulart, and J. Lygeros, "On the road between robust optimization and the scenario approach for chance constrained optimization problems," *IEEE Transactions on Automatic Control, to appear*, 2014. [Online]. Available: http://control.ee.ethz.ch/index.cgi?page=publications&action=details&id=4259

[17] S. Grammatico, X. Zhang, K. Margellos, P. Goulart, and J. Lygeros, "A scenario approach to non-convex control design," *Technical Report, ETH Zürich*, 2013. [Online]. Available: http://control.ee.ethz.ch/~gsergio/GraZhaMarGouLyg_TAC13.pdf

[18] M. Campi and S. Garatti, "A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality," *Journal of Optimization Theory and Applications*, vol. 148, no. 2, pp. 257–280, 2011.

[19] G. Schildbach, L. Fagiano, and M. Morari, "Randomized Solutions to Convex Programs with Multiple Chance Constraints," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2479 – 2501, 2013.

[20] L. Deori, S. Garatti, and M. Prandini, "Stochastic constrained control: trading performance for state constraint feasibility," *Proceeding of European Control Conference*, pp. 2740–2745, 2013.

[21] T. Alamo, R. Tempo, and A. Luque, "On the Sample Complexity of Randomized Approaches to the Analysis and Design under Uncertainty," *American Control Conference*, pp. 4671 – 4676, 2010.