# On the connection between compression learning and scenario based single-stage and cascading optimization problems

Kostas Margellos, Maria Prandini and John Lygeros

***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 provide novel learnability conditions for the underlying algorithms. We then consider different constrained optimization problems affected by uncertainty represented by means of scenarios. We show that the compression learning perspective provides a unifying framework for scenario based optimization, since the issue of providing guarantees on the probability of constraint violation reduces to a learning problem for an appropriately chosen algorithm that satisfies some consistency assumption. To illustrate this, we revisit the scenario approach within the developed context. Moreover, using the compression learning machinery we provide novel results on the probability of constraint violation for the class of cascading optimization problems.**

***Index Terms*—Compression learning, consistent algorithms, randomized optimization, scenario approach, statistical learning theory.**

## I. 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 such problems are proposed under certain assumptions on the for the underlying probability distribution and on the dependency of the constraints on the uncertainty.

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, which involves solving an optimization problem whose constraints depend only on a finite number of uncertainty instances called "scenarios". It does not require any specific assumption on the probability distribution of the uncertainty neither on the way in which the uncertainty enters the problem, but generalizes the properties of the solution to unseen uncertainty instances, providing guarantees on the probability of constraint satisfaction. For problems that are convex with respect to the decision variables the so called scenario approach [7–9] and subsequent contributions [10], [11], offers an already mature theoretical framework for analyzing the generalization properties of the optimal solution. In the non-convex case, tools from statistical learning [12–14] based on the Vapnik-Chervonenkis (VC) theory offer guarantees on the probability that any feasible solution of a scenario based optimization problem satisfies the constraints of the original program [15–17].

In this paper we explore the links between learning theory and the scenario approach [7–9], without resorting to VC theoretic results. To this end we exploit the results of [1] and consider compression learning algorithms, which are based on an alternative notion of learning under an assumption referred to as consistency. A formal definition of consistency will be given in the next section; roughly speaking it refers to the empirical agreement between a set that we seek to learn and our estimate for this set. Our contributions are threefold: 1) We first show how using ideas from the scenario approach theory one can strengthen or relax the consistency assumption which is at the basis of the learning algorithms in [1]. This allows us to extend the key theorem (Theorem 6) of [1] and provide novel learnability conditions (Theorems 3, 4) for a general class of algorithms, not necessarily related to scenario based optimization. 2) The compression learning perspective provides a unifying framework for scenario based optimization since it reveals sufficient conditions for providing guarantees on the probability of constraint satisfaction. In particular, we show that the latter can be equivalently thought of as a learning problem for an appropriately chosen algorithm. In this context we revisit the scenario approach [7–9] and show how the existing probabilistic feasibility bounds follow from our compression learning results. 3) Using the compression learning machinery we address the problem of providing guarantees on the probability of constraint satisfaction for the class of cascading optimization problems. Such problems arise in different contexts, yet, to the best of our knowledge, providing probabilistic bounds on the feasibility of the system constraints has proven to be elusive (e.g. [18]).

Section II introduces the notion of compression and provides certain learnability conditions. In Section III the learning theoretic results are related to scenario based optimization. Section IV deals with cascading optimization and Section V provides some concluding remarks. All omitted proofs can be found in [19].

## II. Learning results

### A. Compression learning

We start by describing some learning concepts and results from [1]. We consider problems affected by an uncertain parameter $\delta$, which is a vector of $n_\delta$ elements, taking values in some set $\Delta \subseteq \mathbb{R}^{n_\delta}$, endowed with a $\sigma$-algebra $\mathcal{D}$. Let $\mathbb{P}$ be a probability measure defined over $\mathcal{D}$. Throughout the paper we impose the following assumption.

**Assumption 1.** *For $m \in \mathbb{N}$, let $\{\delta_i\}_{i=1}^m$ be a collection of $m$ samples $\delta_i \in \Delta$ extracted according to $\mathbb{P}$. Assume that all samples are i.i.d.*

We refer to $\{\delta_i\}_{i=1}^m$ as an $m$-multisample. For any $C \in \mathcal{D}$ 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* set for which we assume that an oracle is available, providing the labeling $\mathbb{1}_T(\delta)$ for any $\delta \in \Delta$. In Section III we consider as target set the entire uncertainty space; this is a case where $T$ may be unknown and only historical data of the uncertainty may be available. The following basic definitions are adapted from [15], where elements of $\mathcal{D}$ are referred to as *concepts*.

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

The labeled multisample is a description of the possibly unknown target set and contains pairs of samples and labels, where the label dictates whether the corresponding sample belongs to the target set.

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 reference [1] to our attention.

K. Margellos is with the Department of Industrial Engineering and Operations Research, UC Berkeley, Sutardja Dai Hall 330, Berkeley CA 94720, United States, e-mail: kostas.margellos@berkeley.edu.

M. Prandini is with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano 20133, Italy, e-mail: prandini@elet.polimi.it.

J. Lygeros is with the Department of Information Technology and Electrical Engineering, ETH Zürich, Physikstrasse 3, Zürich 8092, Switzerland, e-mail: lygeros@control.ee.ethz.ch.

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

Definition 2 implies that $H$ is a consistent hypothesis if it provides the same labeling of $\delta_i$, $i = 1, \ldots, m$, as the target set $T$. The error of $H$ as an approximation of the target set $T$ 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 (Chapter 2.2.2 of [15]) of $T$ and $H$, i.e. $d_{\mathbb{P}}(T, H) = \mathbb{P}\big(\delta \in \Delta : \ \mathbb{1}_H(\delta) \neq \mathbb{1}_T(\delta)\big)$, where $d_{\mathbb{P}}(\cdot, \cdot)$ takes as arguments two sets and returns a probability.

**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 set $T$ 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 $\Delta^m$ 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.** *Let $T \in \mathcal{D}$ be a target set. 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 $\epsilon \in (0,1)$, $m \geq m_0$,*

$$\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_m) > \epsilon\Big\} \leq q(m, \epsilon), \qquad (1)$$

*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 set $T$ (PAC-T).*

The statement of Definition 4 is clearly related to PAC learnability [15] (p. 56), where some $\mathcal{C} \subseteq \mathcal{D}$ is considered and an algorithm is said to be PAC for $\mathcal{C}$ if (1) holds uniformly over target sets $T \in \mathcal{C}$. Here we restrict attention to a specific target set in view of the analysis of Section III. For details the reader is referred to [15], [13].

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 set. 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[1] $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}\big)$ 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 $m \geq d$*

$$\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ \text{there exists } I_d \in \mathcal{I}_d \text{ such that } H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}. \qquad (2)$$

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 (i.e. it agrees with the target set on the $m$-multisample) but the approximation error exceeds $\epsilon$ is low. On the other hand, as $m \to d$ the statement of the theorem is trivial and implies that the left-hand side of (2) tends to one, i.e. if we use all samples to construct the hypothesis, then consistency with respect to the labeled multisample does not posses any generalization properties. This theorem was stated in [1] in the context of sample compression, where the map $G_d$ is referred to as the compression function.

[1]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 $G_d$ when fed with $\{\delta_i\}_{i \in I_d}$.

**Assumption 2.** *Let $T \in \mathcal{D}$ be a target set. Assume that there exists $d$ and $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ 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 $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ with $m \geq d$, there exists $I_d \in \mathcal{I}_d$ such that $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$.*

The second part of Assumption 2 is an empirical generalization statement, since a hypothesis constructed using only $d$ samples is consistent with the entire $m$-multisample. Its first part is trivially satisfied for the optimization problems of the next section. 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 $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in m_d}\big)$. We often refer to the set $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ without its argument. It will always be clear from the context whether $m_d$ refers to the underlying map or to the set of indices $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$. We then have the following theorem, which is stated in [1] without a proof; we provide the proof in the appendix.

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

### *B. Strengthening the consistency assumption*

We show how Assumption 2 can be strengthened, allowing us to tighten the bound in Theorem 2. Our analysis builds on [8], [9], and enables us to extend the learning theoretic results of [1].

**Assumption 3.** *Let $T \in \mathcal{D}$ be a target set. Assume that there exists $d$ and $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ 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 $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ with $m \geq d$, there exists a unique $I_d \in \mathcal{I}_d$ such that $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$.*

The addition over Assumption 2 is that the set $I_d \in \mathcal{I}_d$ for which the requirements of Assumption 3 are satisfied is unique. Define $m_d$, $\big\{A_m\big\}_{m \geq d}$ as in Section II-A and note that, under Assumption 3, $m_d : [\Delta \times \{0,1\}]^m \to \mathcal{I}_d$ is uniquely defined in this case.

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

Theorem 3 constitutes a tighter version of Theorem 2 and the result holds with equality for problems that satisfy Assumption 3. The proof of Theorem 3 is similar to the second part of the proof of Theorem 1 in [8], and relies on the fact that, under Assumption 3, the sets $\big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ H_{I_d}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big\}$, for all $I_d \in \mathcal{I}_d$ with $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}\big)$ form a partition of $\Delta^m$ up to a set of measure zero (see [19] for more details).

### *C. Relaxing the consistency assumption*

We now relax Assumption 2 and study its effect on Theorem 2. Motivated by [20], [9], we show how "sampling-and-discarding"

ideas from scenario based optimization can be incorporated in the set-up of Section II-A, enriching the analysis of [1]. Fix $r,\ d \in \mathbb{N}$ and consider $m \geq d+r$. Given a set $I_r \in \mathcal{I}_r$, let $\mathcal{I}_d^{m-r}$ with cardinality $\binom{m-r}{d}$ contain all sets $I_d$ with $d$ indices from $\{1,\dots,m\} \setminus I_r$.

**Assumption 4.** *Let $T \in \mathcal{D}$ be a target set. Assume that there exists $d$ and $G_d : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ 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 $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ 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 $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big)$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\dots,m\}\setminus I_r}$.*
*3) With $\mathbb{P}^m$-probability one, for any $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ 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 two parts of the assumption, $H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big)$ is not consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}$, for all $i \in I_r$.*

The relaxation compared to Assumption 2 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 4 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 $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\dots,m\}\setminus\bar{I}_r}$ and is not consistent with $\big(\delta_i, \mathbb{1}_T(\delta_i)\big)$, for all $i \in \bar{I}_r$. Construct $\big\{A_m\big\}_{m\geq d+r}$, 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 set and fix $r \in \mathbb{N}$. Under Assumption 4, $\big\{A_m\big\}_{m\geq d+r}$ is PAC-T with $q(m,\epsilon) = \binom{m}{d}\sum_{i=0}^r \binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i}$.*

The proof of Theorem 4 follows closely the proof of Theorem 2.1 in [20], but it is not concerned with optimization problems. It requires also modifying the last part of the proof of Theorem 2.1 in [20], which involves integration by parts of a quantity that depends on the bound of Theorem 2, instead of the one of Theorem 3. We can strengthen Assumption 4 by requiring the set $I_d \in \mathcal{I}_d^{m-r}$ that satisfies its requirements to be unique. Theorem 4 holds then with $q(m,\epsilon) = \binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}$. This result is then identical to the one obtained in an optimization context in [20]. The possibility of learning while being inconsistent with a certain fraction of the samples is mentioned in [1], but to the best of our knowledge no specific results in this direction have been published.

## III. Connection to optimization

### A. Scenario based optimization as a learning problem

We show how scenario based optimization can be thought of as a learning problem in the sense of Section II-A. To this end consider the robust optimization problem

$$\mathcal{P} : \min_{x\in\mathcal{X}} c^T x \text{ subject to: } g(x,\delta) \leq 0,\ \forall \delta \in \Delta, \tag{3}$$

where $\mathcal{X} \subset \mathbb{R}^{n_x}$, $c \in \mathbb{R}^{n_x}$ and $g : \mathcal{X}\times\Delta \to \mathbb{R}$. Note that $n_x$ denotes the number of elements of the vector $x \in \mathcal{X}$. As in Section II, assume that $\Delta$ is endowed with a $\sigma$-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 $g(x,\delta) = \max_{j=1,\dots,n_c} g_j(x,\delta)$. Considering a linear objective function is also without loss of generality; in case of a generic objective function, an epigraphic reformulation can be employed [7].

Problem $\mathcal{P}$ is generally difficult to solve when $\Delta$ is a continuous set. We replace $\Delta$ by the discrete set $\{\delta_i\}_{i=1}^m \in \Delta^m$, where the $m$ samples are extracted i.i.d according to $\mathbb{P}$.

$$\mathcal{P}[\{\delta_i\}_{i=1}^m] : \min_{x\in\mathcal{X}} c^T x \text{ subject to: } g(x,\delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m. \tag{4}$$

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

$$H_{I_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big) = \big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in I_d}),\delta) \leq 0\big\}. \tag{5}$$

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. $d_{\mathbb{P}}(T,H_{I_d}) = \mathbb{P}(\{\delta \in \Delta :\ \delta \notin H_{I_d}\}) = \mathbb{P}\big(\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in I_d}),\delta) > 0\}\big)$. Suppose that $d$, $G_d$ are such that Assumption 2 is satisfied. Then there exists a set of indices $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) \in \mathcal{I}_d$ such that $H_{m_d} = \big\{\delta \in \Delta :\ g(x_d(\{\delta_i\}_{i\in m_d}),\delta) \leq 0\big\}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Note that Assumption 2 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 $\mathcal{P}[\{\delta_i\}_{i=1}^m]$.

**Lemma 1.** *Let $T = \Delta$ be the target set and consider Assumption 2. 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 $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) \in \mathcal{I}_d$ that satisfies the second part of Assumption 2. Then, for any $\epsilon \in (0,1)$, $\mathbb{P}^m\big\{(\delta_1,\dots,\delta_m) \in \Delta^m :\ \mathbb{P}\big(\delta \in \Delta :\ g(x_m(\{\delta_i\}_{i=1}^m),\delta) > 0\big) > \epsilon\big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}$.*

Lemma 1 shows that under Assumption 2, 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 feasibility guarantees. With probability at least $1 - \binom{m}{d}(1-\epsilon)^{m-d}$, $x_m$ satisfies (3) except for a set with $\mathbb{P}$-measure at most $\epsilon$. The proof of Lemma 1 is based on showing that an algorithm is PAC-T for $T = \Delta$. This algorithm can be constructed as $\big\{A_m\big\}_{m\geq d}$, where $A_m : [\Delta\times\{0,1\}]^m \to \mathcal{D}$ is such that $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ and $H_m = H_{m_d}$, where $H_m = \big\{\delta \in \Delta :\ g(x_m(\{\delta_i\}_{i=1}^m),\delta) \leq 0\big\}$. $H_m = H_{m_d}$ is equivalent to $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in m_d})$. The latter is satisfied in the set-up of Section III-B and the other cases in [19].

Replacing Assumption 2 with Assumption 3, Lemma 1 remains valid with its bound replaced by the one of Theorem 3; in fact the result will hold with equality. One can also relax Assumption 2 (see discussion at the end of Section II-C) such that the bound of Lemma 1 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 multisample 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 [20] and as constraint removal in [9].

### B. The scenario approach

We next consider the set-up of the scenario approach as proposed in [7]. We show that by appropriately selecting the constraint functions of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and the map $x_m : \Delta^m \to \mathcal{X}$, Assumption 2 is satisfied, obtaining feasibility guarantees by virtue of Lemma 1.

**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, the feasibility region $\big\{x \in \mathcal{X} :\ g(x,\delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\}$ of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ has a non-empty interior and the minimizer of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ 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 $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and note that $x_m$ belongs to the feasibility region of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$. The scenario approach is based on the notion of support constraints. A constraint in $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ 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 [10], [11], an improved bound based on the dimension of the unconstrained decision space is provided. The subsequent analysis is valid for any upper bound on the number of support constraints. Therefore, let the number of support constraints be *at most* $\zeta < \infty$.

**Lemma 2.** *Let $T = \Delta$ be the target set 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 $\mathcal{P}[\{\delta_i\}_{i\in I_d}]$. $G_d$ then satisfies Assumption 2.*

Under Lemma 2, there exists $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) \in \mathcal{I}_d$ with $d = \zeta$ such that $H_{m_d} = \big\{\delta \in \Delta : \ g(x_d(\{\delta_i\}_{i\in m_d}), \delta) \leq 0\big\}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. As shown in the proof of Lemma 2, the set $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ for which Assumption 2 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 $\mathcal{P}[\{\delta_i\}_{i=1}^m]$. Therefore, the bound of Lemma 1 holds. Such a conclusion is identical to Theorem 1 of [7] (with $n_x$ in place of $\zeta$).

An improved bound is given in Theorem 1 of [8]. To recover this, in addition to Assumption 5 assume that $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ is such that, with $\mathbb{P}^m$-probability one, the number of support constraints is *equal to* $\zeta$. The following cases can then be distinguished: 1) In the particular case where $d = \zeta = n_x$, we have the class of fully supported problems [8]. Considering problems where $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ has exactly $\zeta$ support constraints is a sufficient condition for Assumption 3 to be satisfied (see Proposition 4 in [19]). In that case the bound of Lemma 1 can be replaced by the bound of Theorem 3. Moreover, the result would be tight and would hold with equality. 2) If the problem does not have exactly $\zeta$ support constraints almost surely, we can still obtain similar probabilistic guarantees following [9], [8]. Specifically, if a problem is non-degenerate and has at most $\zeta$ support constraints, then by a procedure called regularization [9], [10], it can be transformed to a different problem with exactly $\zeta$ support constraints. One can then bound the probability in the left-hand side of the bound of Lemma 1 by the probability of constraint violation for the regularized problem, which is equal to $\sum_{i=0}^{\zeta-1} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i}$. 3) If the problem does not have exactly $\zeta$ support constraints almost surely, but is degenerate, the aforementioned bound is still valid, as shown in [8] using a "heating-cooling" procedure.

## IV. Cascading optimization problems

We consider here the class of cascading optimization problems and show how we can employ the learning theoretic machinery of Section II-A to obtain guarantees on the probability of satisfying the constraints in all problems in the cascade. Every problem in the cascade is a scenario program that depends 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 on the probability of simultaneous satisfaction of the constraints of all 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 extend any finite number of cascading problems.

For any $m \in \mathbb{N}$, consider the following family of problems:

$$\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i=1}^m] : \min_{y\in\mathcal{Y}} \tilde{c}^T y \text{ subject to: } \widetilde{g}(y, x, \delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m, \quad (6)$$

which is parametric in the vector of decision variables $x \in \mathcal{X}$ of an optimization problem of the form of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ in (4), $\mathcal{Y} \subset \mathbb{R}^{n_y}$, $\tilde{c} \in \mathbb{R}^{n_y}$, and $\widetilde{g} : \ \mathcal{Y} \times \mathcal{X} \times \Delta \to \mathbb{R}$. Note that $n_y$ denotes the number of elements of the vector of decision variables $y \in \mathcal{Y}$.

**Assumption 6.** *Suppose that $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ satisfies Assumption 5. 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.*

Under Assumption 6, Lemma 2 implies that Assumption 2 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 the hypothesis $H_{I_{d_1}} = \big\{\delta \in \Delta : \ g(x_{d_1}(\{\delta_i\}_{i\in I_{d_1}}), \delta) \leq 0\big\}$, 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 2 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 $H_{I_{d_2}}[x] = \widetilde{G}_{d_2}[x]\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_{d_2}}\big) = \big\{\delta \in \Delta : \ \widetilde{g}(y_{d_2}[x](\{\delta_i\}_{i\in I_{d_2}}), x, \delta) \leq 0\big\}$, where $y_{d_2}[x] : \Delta^{d_2} \to \mathcal{Y}$ is the unique, under Assumption 6, minimizer of $\widetilde{\mathcal{P}}[x, \{\delta_i\}_{i\in I_{d_2}}]$.

**Lemma 3.** *Let $T = \Delta$ be the target set and consider Assumption 6. Fix $d = d_1 + d_2$ and consider $m \geq d$. For any $I_d \in \mathcal{I}_d$, construct $G_d^c : [\Delta \times \{0,1\}]^d \to \mathcal{D}$ as in (7). $G_d^c$ then satisfies Assumption 2.*

$$\begin{aligned} &G_d^c\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_d}\big) = H_{I_d} \cap \widetilde{H}_{I_d}[x_d(\{\delta_i\}_{i\in I_d})] \\ &= \big\{\delta \in \Delta : \ \big(g(x_d(\{\delta_i\}_{i\in I_d}), \delta) \leq 0\big) \\ &\qquad \text{and } \big(\widetilde{g}(y_d[x_d(\{\delta_i\}_{i\in I_d})](\{\delta_i\}_{i\in I_d}), x_d(\{\delta_i\}_{i\in I_d}), \delta) \leq 0\big)\big\}. \end{aligned} \quad (7)$$

Lemma 3 shows that if there exist a compression function for two optimization problems, then there exists a compression function for the cascade of these problems. Under Lemma 3, there exists $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) \in \mathcal{I}_d$ 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 5.** *Let $T = \Delta$ be the target set and consider Assumption 6. Fix $d = d_1 + d_2$ and consider $m \geq d$. Then, for any $\epsilon \in (0,1)$,*

$$\begin{aligned} &\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ \mathbb{P}\Big(\delta \in \Delta : \ \big(g(x_m, \delta) > 0\big) \\ &\qquad \text{or } \big(\widetilde{g}(y_m[x_m], x_m, \delta) > 0\big)\Big) > \epsilon\Big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}, \end{aligned} \quad (8)$$

*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.*

Note that, in a scenario approach context, $d$ is the sum of the number of support constraints of each problem in the cascade. Theorem 5 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 set $T = \Delta$. This algorithm comprises $A_m : [\Delta \times \{0,1\}]^m \to \mathcal{D}$ such that $H_m = A_m\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{j=1}^m\big)$ 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 Lemma 2. We refer to $\{A_m\}_{m\geq d}$ as cascading algorithm

since it is constructed based on a cascade of two sequentially dependent hypotheses. We only need to invoke Assumption 6 in the proof of Lemma 3 and Theorem 5, where a by-product of Lemma 2 is employed. In [19] we discuss how this assumption can be relaxed.

Note that, under Assumption 6, 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 $F = \big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ \forall x \in \big\{x \in \mathcal{X} : \ g(x, \delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\}, \big\{y \in \mathcal{Y} : \ \widetilde{g}(y, x, \delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\} \neq \emptyset\big\}$. $F$ is a restriction of $\Delta^m$ 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 $x \in \big\{x \in \mathcal{X} : \ g(x, \delta) \leq 0, \forall \delta \in \{\delta_i\}_{i=1}^m\big\}$, i.e. for any $x$ for which the first problem in the cascade is feasible. The result of Theorem 5 will then still hold if we replace $\Delta^m$ with $F$ in (8).

Theorem 5 implies that the solution comprising the solutions of the individual problems in the cascade is feasible for the constraints of both problems. It follows then directly from (8) that the probability of constraint violation for each of the two problems is also bounded by the quantity on the right-hand side of (8). Note that the second problem in the cascade is allowed to have an arbitrary dependence on $x$ (see Assumption 6). One example of a problem with constraint functions that are not jointly convex with respect to $x$ 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 is to follow an iterative algorithm with an a-priori fixed number of iterations, alternating between optimization problems that involve either $x$ or $y$, having the other decision vector fixed to the value obtained at the preceding iteration. Alternatively, since the problem is non-convex, to provide guarantees in the form of (8) 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 [18], where a cascade of two scenario programs was formulated to address the multi-objective nature of the problem, but the violation properties of the resulting solution were not discussed. The first problem in the cascade was in the form of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ (satisfying Assumption 5) with the constraint function encoding the input constraints. At the second problem in the cascade, a bound on the system sate was introduced and was considered as a decision variable. The objective was to minimize this (soft) bound, subject to both input and state constraints 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 $\mathcal{P}[\{\delta_i\}_{i=1}^m]$, and it is necessary to use the same samples with the first one to ensure feasibility. This two-step approach 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 $\alpha$ compared to the value obtained at the first problem. This set-up fits our cascading framework with $n_y = n_x + 1$ (the additional variable is due to the soft bound) and $F = \Delta^m$. Using the same samples for both problems in the cascade is not only crucial for feasibility purposes. In bilinear descent type of algorithms, using the same samples at every problem in the cascade, the objective function is confined to decrease at every iteration of the algorithm. For more applications and comparison with other scenario based implementations the reader is referred to [21].

Unfortunately, for cascading problems we cannot provide the tighter bound of Theorem 3. Even if we replace Assumption 2 with Assumption 3 in Lemma 3, 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 Lemma 3, satisfies Assumption 3 (see also the construction of a set $I_d$ that satisfies Assumption 2 in the proof of Lemma 3). However, one can relax Assumption 2 in Theorem 5 to Assumption 4 and replace the right-hand side of (8) according to Theorem 4. To ensure that the obtained solution violates the removed constraints, thus satisfying the last part of Assumption 4, 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 eliminate a sample only if it results in a reduction in the objective value of the subproblem of interest.

## V. Concluding remarks

We considered a compression learning paradigm for algorithms that satisfy some consistency assumption. It was shown how using results from the scenario approach we can strengthen or relax this assumption, providing novel learnability conditions for a general class of algorithms, not necessarily related to optimization. Concentrated on scenario based optimization problems we then showed that guarantees on the probability of constraint violation can be provided by treating them as learning problems. We also showed how one can exploit the developed machinery to provide guarantees on the probability of constraint satisfaction for the class of cascading optimization problems. These novel results demonstrate how compression learning can prove useful for scenario based multi-objective and sequential optimization problems. Our developments extend also to other cases, like those in [22]. Details can be found in [19].

## Appendix

**Proof of Theorem 2.** Consider any $\epsilon \in (0, 1)$. Under Assumption 2, let $m_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big) \in \mathcal{I}_d$ 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 $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. Then, $\mathbb{P}^m\big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_{m_d}) > \epsilon\big\} = \mathbb{P}^m\big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ H_{m_d}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ and $d_{\mathbb{P}}(T, H_{m_d}) > \epsilon\big\}$. Now since the last term is upper bounded by $\mathbb{P}^m\big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ \exists I_d \in \mathcal{I}_d$ such that $H_{I_d}$ is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$ and $d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\big\}$, by Theorem 1, we have that $\mathbb{P}^m\big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_{m_d}) > \epsilon\big\} \leq \binom{m}{d}(1-\epsilon)^{m-d}$. Set $q(m, \epsilon) = \binom{m}{d}(1-\epsilon)^{m-d}$. Since $\binom{m}{d} \leq \sum_{i=0}^{d}\binom{m}{i} \leq \big(\frac{me}{d}\big)^d$ (the second inequality is due to Lemma 4.3 of [15]), we have that $\lim_{m\to\infty} q(m, \epsilon) \leq \lim_{m\to\infty}\big(\frac{me}{d}\big)^d(1-\epsilon)^{m-d} = 0$. Therefore, $\lim_{m\to\infty} q(m, \epsilon) = 0$. Construct then algorithm $\big\{A_m\big\}_{m \geq d}$, where $A_m : [\Delta \times \{0, 1\}]^m \to \mathcal{D}$ 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 Lemma 1.** Under Assumption 2, $H_{m_d} = \big\{\delta \in \Delta : \ g(x_d(\{\delta_i\}_{i \in m_d}), \delta) \leq 0\big\}$ 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 $\big\{A_m\big\}_{m \geq d}$, 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 $H_m = \big\{\delta \in \Delta : \ g(x_m(\{\delta_i\}_{i=1}^m), \delta) \leq 0\big\}$. Moreover, by the assumption of the lemma $x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i \in m_d})$, which entails that $H_m = H_{m_d} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in m_d}\big)$, for $G_d$ defined

according to (5). 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 $T=\Delta$, $d_{\mathbb{P}}(T,H_m)=\mathbb{P}\big(\{\delta\in\Delta:\ g(x_m(\{\delta_i\}_{i=1}^m),\delta)>0\}\big)$, concludes the proof. □

**Proof of Lemma 2.** 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 $m_d\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m\big)\in\mathcal{I}_d$ 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 $\mathcal{P}[\{\delta_i\}_{i=1}^m]$ and $\mathcal{P}[\{\delta_i\}_{i\in m_d}]$, respectively. The solution $x_d(\{\delta_i\}_{i\in m_d})$ satisfies all constraints that correspond to samples whose indices are not included in $m_d\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m\big)$, 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 $i\in\{1,\dots,m\}\setminus m_d$. But, since $x_d(\{\delta_i\}_{i\in m_d})$ is the optimal solution of $\mathcal{P}[\{\delta_i\}_{i\in m_d}]$ 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 $i\in\{1,\dots,m\}$ and since $H_{m_d}=\big\{\delta\in\Delta:\ g(x_d(\{\delta_i\}_{i\in m_d}),\delta)\leq 0\big\}$, we have that $\mathbb{1}_{H_{m_d}}(\delta_i)=1$, for all $i=1,\dots,m$. The last statement, together with the fact that $T=\Delta$, implies that $H_{m_d}=G_d\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i\in m_d}\big)$ is consistent with $\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m$, showing the second part of Assumption 2.

It remains to show the first part of Assumption 2. For any $I_d\in\mathcal{I}_d$, since $x_d(\{\delta_i\}_{i\in I_d})$ is the minimizer of $\mathcal{P}[\{\delta_i\}_{i\in I_d}]$ 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 Lemma 3.** Under Assumption 5, Assumption 2 holds for $d_1\in\mathbb{N}$, $G_{d_1}:[\Delta\times\{0,1\}]^{d_1}\to\mathcal{D}$, i.e. $\exists\, m_{d_1}(\{\delta_i\}_{i=1}^m)\in\mathcal{I}_{d_1}$: $H_{m_{d_1}}=G_{d_1}\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i\in m_{d_1}}\big)$ is consistent with $\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m$. Since $H_{m_{d_1}}=\big\{\delta\in\Delta:\ g\big(x_{d_1}(\{\delta_i\}_{i\in m_{d_1}}),\delta\big)\leq 0\big\}$,

$$g\big(x_{d_1}(\{\delta_i\}_{i\in m_{d_1}}),\delta_i\big)\leq 0,\ \text{for all } i=1,\dots,m. \tag{9}$$

Moreover, under Assumption 6, for all $x\in\mathcal{X}$, Assumption 2 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 $\widetilde{H}_{m_{d_2}[x]}[x]=\widetilde{G}_{d_2}[x]\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i\in m_{d_2}[x]}\big)$ is consistent with $\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m$. 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 $x\in\mathcal{X}$,

$$\widetilde{g}\big(y_{d_2}[x](\{\delta_i\}_{i\in m_{d_2}[x]},x,\delta_i)\big)\leq 0,\ \text{for all } i=1,\dots,m. \tag{10}$$

Set $d=d_1+d_2$ and consider $m\geq d$. Choose a set of indices $m_d\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m\big)\in\mathcal{I}_d$ such that $m_d\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m\big)\supseteq m_{d_1}\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m\big)\cup m_{d_2}[x_{d_1}(\{\delta_i\}_{i\in m_{d_1}(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m)})]\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m\big)$. 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. Recall that $x_{d_1}(\{\delta_i\}_{i\in m_{d_1}})$ is the minimizer of $\mathcal{P}[\{\delta_i\}_{i\in m_{d_1}}]$ used to construct $H_{m_{d_1}}$. For simplicity, 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 Lemma 2, 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, (9) implies that $g\big(x_d(\{\delta_i\}_{i\in m_d}),\delta_i\big)\leq 0$, for all $i=1,\dots,m$. We also have that $m_d\supseteq m_{d_2}[x_{d_1}(\{\delta_i\}_{i\in m_{d_1}})]=m_{d_2}[x_d(\{\delta_i\}_{i\in m_d})]$, with the equality due to 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 Lemma 2, $y_d[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_d})=y_{d_2}[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_{d_2}[x_d(\{\delta_i\}_{i\in m_d})]})$. By (10) we then have that $\widetilde{g}\big(y_d[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_d}),x_d(\{\delta_i\}_{i\in m_d}),\delta_i)\big)\leq 0$, for all $i=1,\dots,m$. Therefore, for all $i=1,\dots,m$,

$$\begin{aligned}&g\big(x_d(\{\delta_i\}_{i\in m_d}),\delta_i\big)\leq 0\\ &\text{and } \widetilde{g}\big(y_d[x_d(\{\delta_i\}_{i\in m_d})](\{\delta_i\}_{i\in m_d}),x_d(\{\delta_i\}_{i\in m_d}),\delta_i)\big)\leq 0.\end{aligned} \tag{11}$$

Since $T=\Delta$, (11), (7) imply that $G_d^c\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i\in m_d}\big)=H_{m_d}\cap\widetilde{H}_{m_d}[x_d(\{\delta_i\}_{i\in m_d})]$ is consistent with $\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m$. To conclude the proof it remains to show the first part of Assumption 2; this is done as in the proof of Lemma 2. □

**Proof of Theorem 5.** Under Assumption 6, Lemma 3 implies that $G_d^c$ satisfies Assumption 2. Then, there exists $m_d\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m\big)\in\mathcal{I}_d$ such that $H_{m_d}^c=\big\{\delta\in\Delta:\ \big(g(x_{m_d},\delta)\leq 0\big)$ and $\big(\widetilde{g}(y_{m_d}[x_{m_d}],x_{m_d},\delta)\leq 0\big)\big\}$ is consistent with $\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i=1}^m$. 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 $H_m=\big\{\delta\in\Delta:\ \big(g(x_m,\delta)\leq 0\big)$ and $\big(\widetilde{g}(y_m[x_m],x_m,\delta)\leq 0\big)\big\}$. Under Assumption 6, from the proof of Lemma 2 we have $x_m(\{\delta_i\}_{i=1}^m)=x_d(\{\delta_i\}_{i\in m_d})$, $y_m[x_m](\{\delta_i\}_{i=1}^m)=y_d[x_d](\{\delta_i\}_{i\in m_d})$ and hence $H_m=H_{m_d}^c=G_d^c\big(\{(\delta_i,\mathbb{1}_T(\delta_i))\}_{i\in m_d}\big)$. By Theorem 2, $\{A_m\}_{m\geq d}$ is then PAC-T with $q(m,\epsilon)=\binom{m}{d}(1-\epsilon)^{m-d}$. The latter, together with the fact that, since $T=\Delta$, $d_{\mathbb{P}}(T,H_m)=\mathbb{P}\big(\delta\in\Delta:\ \big(g(x_m,\delta)>0\big)$ or $\big(\widetilde{g}(y_m[x_m],x_m,\delta)>0\big)\big)$, leads to (8). □

## References


[1] S. Floyd and M. Warmuth, "Sample compression, learnability, and the Vapnik-Chervonenkis dimension," *Mach. Learn.*, 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 ," *Math. Progr.*, vol. 112, pp. 183 – 183, 2008.

[5] A. Nemirovski and A. Shapiro, "Convex Approximations of Chance Constrained Programs," *SIAM J. Control Optimiz.*, vol. 17, no. 4, pp. 969 – 996, 2006.

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

[7] G. Calafiore and M. Campi, "The scenario approach to robust control design," *IEEE Trans. Autom. 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 J. Optimiz.*, vol. 19, no. 3, pp. 1211–1230, 2008.

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

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

[11] G. Schildbach, L. Fagiano, C. Frei, and M. Morari, "The Scenario Approach for Stochastic Model Predictive Control with Bounds on Closed-Loop Constraint Violations," *Automatica*, vol. 50, no. 12, pp. 3009–3018, 2014.

[12] 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.

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

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

[15] M. Vidyasagar, *A Theory of Learning and Generalization*. London: Springer, second edition, 2002.

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

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

[18] L. Deori, S. Garatti, and M. Prandini, "Stochastic constrained control: trading performance for state constraint feasibility," *Eur. Control Conf.*, pp. 2740–2745, 2013.

[19] K. Margellos, M. Prandini, and J. Lygeros, "On the connection between compression learning and scenario based optimization," *Tech. Rep.*, pp. 1–29, 2014. [Online]. Available: http://arxiv.org/abs/1403.0950

[20] M. Campi and S. Garatti, "A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality," *J. Opt. Theory and Appl.*, vol. 148, no. 2, pp. 257–280, 2011.

[21] N. Kariotoglou, K. Margellos, and J. Lygeros, "On the computational complexity and generalization properties of multi-stage and recursive scenario programs," *IEEE Trans. Autom. Control (under review)*, pp. 1–15, 2014. [Online]. Available: http://arxiv.org/pdf/1412.4203.pdf

[22] K. Margellos, P. Goulart, and J. Lygeros, "On the road between robust optimization and the scenario approach for chance constrained optimization problems," *IEEE Trans. Autom. Control*, vol. 59, no. 8, pp. 2258 – 2263, 2014.