Margin-based scenario approach to robust optimization in high dimension

Fabien Lauer

Margin-based scenario approach to robust optimization in high dimension

Fabien Lauer

***Abstract*—This paper deals with the scenario approach to robust optimization. This relies on a random sampling of the possibly infinite number of constraints induced by uncertainties in the parameters of an optimization problem. Solving the resulting random program yields a solution for which the quality is measured in terms of the probability of violating the constraints for a random value of the uncertainties, typically unseen before. Another central issue is the determination of the sample complexity, i.e., the number of random constraints (or scenarios) that one must consider in order to guarantee a certain reliability. In this paper, we introduce an additional margin in the constraints and analyze the probability of violation of solutions to the modified random programs. In particular, using tools from statistical learning theory, we show that the sample complexity of a class of problems does not explicitly depend on the number of variables. In addition, within this class, that includes polynomial constraints among others, the same guarantees hold for both convex and nonconvex instances.**

***Index Terms*—Scenario approach, Robust optimization, Randomized algorithms, Statistical learning**

I. Introduction

ROBUST optimization is concerned with parametrized optimization problems with uncertainties on the parameters. The worst-case approach to these problems aims at computing a solution that satisfies the constraints for all possible values of the parameters within the uncertainty set. However, this often leads to an infinite number of constraints that cannot always be handled in practice. Instead, we focus on the scenario approach, in which the uncertainties are sampled to generate a finite number of random constraints. Then, the issue is to determine the probability with which the corresponding solution satisfies the constraints, or, conversely, how many samples should be drawn to guarantee a certain upper bound on the probability of violation.

The scenario approach to robust optimization has a long history [1]–[8], also tightly connected with the field of robust control [9]. This paper derives new results inspired by statistical learning theory, thus following the path initiated by [3]. However, while [3] focused solely on the VC-dimension and early tools from learning theory, we introduce a refined analysis based on more recent developments in this field. More precisely, the approach of [3] focuses on whether the scenario constraints are satisfied or not. Instead, we introduce the concept of margin to measure the amount by which a scenario constraint is satisfied and then analyze the random programs through a capacity measure known as the Rademacher complexity [10], [11], which takes into account the magnitude of the constraint function in addition to its sign. More precisely, we focus on modified versions of the random programs and solutions that satisfy tighter scenario constraints in order to ease the satisfaction of the original constraint in generalization.

F. Lauer is with the Université de Lorraine, LORIA, CNRS, Nancy, France (e-mail: fabien.lauer@loria.fr).

A. Contributions

The main contributions of the paper can be summarized as follows.

1) Dimension-free a priori bounds: We derive bounds on the probability of violation that do not explicitly depend on the dimension d of the optimization problem (the number of variables), and thus show that the sample complexity of a margin-based random problem is not directly related to its dimension, but instead depends on the level of margin with which the scenario constraints are satisfied. This stands in contrast to most of the literature: For convex problems, the sample complexity derived in [2], [4] grows linearly with the dimension, while extensions of these works to nonconvex problems, such as the approach of [5] for integer programming, suffer from the same issue. More precisely, [2], [4] show that the probability of violation can be bounded in terms of the length of the support subsample, but this is a random quantity, itself difficult to bound a priori otherwise than by the dimension for general convex problems. As noted by [12], this can be problematic, e.g., for model predictive control, where the number of variables can grow quadratically with the horizon. Other works [6]–[8] provide bounds on the probability of violation that do not involve d on the basis of the a posteriori length of the support subsample. However, these bounds only apply after the solution has been computed and do not yield a sample complexity. An incremental approach was recently proposed in [13] to estimate the sample complexity on the basis of a posteriori bounds. However, this requires solving multiple optimization programs, and is beneficial mostly in cases where scenarios are more expensive than computations. Instead, we here focus on the computational burden related to the sample complexity for cases where the scenarios can be easily generated.

2) Convex problems: Random convex programs were thoroughly studied in [2], [4]. In particular, a priori bounds on the probability of violation in $O(d/N)$ for problems with d variables and N scenarios were derived. It was also shown that these could not be improved in general. However, by restricting the class of problems to those with a particular structure, it was shown in [12] that tighter bounds could be obtained. In this paper, we follow the latter while enlarging its scope: our proposed method can handle more general constraints than [12] and yields margin-based solutions with improved reliability guarantees compared with those of [2], [4] in the high-dimensional regime.

3) Nonconvex problems: The bounds derived in this paper hold independently of the convexity of the problem. For instance, a convex problem over d real variables can be reformulated over integers without impacting the statistical guarantees. More generally, our framework includes all combinations of conjunctions and disjunctions of polynomial constraints composed with Lipschitz functions without uncertainty in the exponents, over either real or integer variables. In addition, the guarantees derived in this paper hold without assumptions on our ability to solve the random scenario program. At the opposite, approaches that extend the results of [4] to the nonconvex case apply only to the global optimizer [5] or to solutions that at least satisfy all scenario constraints [7].

B. Notations

For two vectors $u,\ v \in \mathbb{R}^n$, the dot product is written as $u^\top v$ and $\|u\|$ denotes the Euclidean norm. Given two sets $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{Y}^\mathcal{X}$ is the set of functions of $\mathcal{X}$ to $\mathcal{Y}$. For two functions $f \in \mathbb{R}^\mathcal{X}$ and $\varphi \in \mathbb{R}^\mathbb{R}$, $\varphi \circ f$ denotes their composition, i.e., $\varphi \circ f(x) = \varphi(f(x))$, and for a class of functions $\mathcal{F} \subset \mathbb{R}^\mathcal{X}$, $\varphi \circ \mathcal{F} = \{\varphi \circ f,\ f \in \mathcal{F}\}$. Given two sets of functions, $\mathcal{U},\ \mathcal{V} \subset \mathbb{R}^\mathcal{X}$, we also introduce natural notations for the sets induced by pointwise binary operations: $\mathcal{U}+\mathcal{V} = \{u+v,\ u \in \mathcal{U},\ v \in \mathcal{V}\}$, $\mathcal{U}-\mathcal{V} = \{u-v,\ u \in \mathcal{U},\ v \in \mathcal{V}\}$, $\max(\mathcal{U},\mathcal{V}) = \{g \in \mathbb{R}^\mathcal{X} : g(x) = \max\{u(x), v(x)\},\ u \in \mathcal{U},\ v \in \mathcal{V}\}$, $\min(\mathcal{U},\mathcal{V}) = \{g \in \mathbb{R}^\mathcal{X} : g(x) = \min\{u(x), v(x)\},\ u \in \mathcal{U},\ v \in \mathcal{V}\}$. The operator $\mathbb{E}_\theta$ denotes the expectation wrt. the random variable θ, and we write merely $\mathbb{E}$ when the random variable is obvious from the context. The indicator function is written as $\mathbf{1}_E$ and is 1 when the expression E is true and 0 otherwise.

C. Paper organization

We start in Section II by formulating the class of problems we consider. Then, after recalling the standard scenario approach in Sect. II-A, we introduce the margin and the proposed approach in Sect. II-B. Section III contains the theoretical results, including the main generalization bounds in Sect. III-A. The sample complexity is discussed in Sect. III-B, and the maximization of the margin in Sect. III-C. Other possibilities are briefly discussed, for a posteriori bounds in Sect. III-D, and bounds with fast rates of convergence in Sect. III-E. Finally, optimization problems are considered in Sect. IV and numerical results are reported in Sect. V before the conclusions in Sect. VI.

II. Margin-based scenario approach

Given the set of admissible parameter values Θ accounting for the uncertainties, the prototype robust optimization problem that we consider is as follows.

Problem 1 (Robust program)**.** *Given a parameter set Θ, a domain $\mathcal{X}$, a cost function $J : \mathcal{X} \to \mathbb{R}$ and a constraint function $f : \mathcal{X} \times \Theta \to \mathbb{R}$, solve*

$$\min_{x \in \mathcal{X}} J(x), \quad \textit{s.t.} \quad f(x,\theta) \leq 0, \quad \forall \theta \in \Theta.$$

Here, the domain $\mathcal{X}$ entails all the constraints that are not impacted by uncertainties. No assumption is made regarding its convexity, and $\mathcal{X}$ can typically be a subset of $\mathbb{R}^d$ or $\mathbb{Z}^d$.

While the proposed approach is generally applicable in principle to any constraint function f, we will also provide dedicated results for problems that can be formulated with

$$f(x,\theta) = \max_{k \in \{1,\dots,C\}} f_k(x,\theta) \tag{1}$$

for a collection of C functions

$$f_k(x,\theta) = \psi_k(\theta)^\top \phi_k(x) + \eta_k(\theta), \quad k = 1,\dots,C, \tag{2}$$

based on functions $\psi_k : \Theta \to \mathbb{R}^{n_k}$, $\phi_k : \mathcal{X} \to \mathbb{R}^{n_k}$ and $\eta_k : \Theta \to \mathbb{R}$. Again, no assumption will be made regarding the convexity of these functions. In fact, the constraint $f(x,\theta) \leq 0$ with the form (1), which implements a conjunction of C constraints (all C functions $f_k(x,\theta)$ must be negative), can be replaced by the nonconvex form

$$f(x,\theta) = \min_{k \in \{1,\dots,C\}} f_k(x,\theta), \tag{3}$$

which implements a disjunction of C alternatives (at least one $f_k(x,\theta)$ must be negative). Any combination of max and min operations can also be considered to implement for instance $f(x,\theta) = \max\{f_3(x,\theta), \min\{f_1(x,\theta), f_2(x,\theta)\}\}$. In full generality, we will thus focus on constraint functions that take the form

$$\begin{aligned} f(x,\theta) = \rho_C \circ g_C(\ \dots \rho_3 \circ g_3(\ \rho_2 \circ g_2(\ \varphi_1 \circ f_1(x,\theta)\ , \\ \varphi_2 \circ f_2(x,\theta)\ ),\ \varphi_3 \circ f_3(x,\theta))\dots\ , \varphi_C \circ f_C(x,\theta)) \end{aligned} \tag{4}$$

for $C-1$ binary operators

$$\begin{aligned} g_k \in \{&(a,b) \mapsto \max(a,b), (a,b) \mapsto \min(a,b), \\ &(a,b) \mapsto a+b, (a,b) \mapsto a-b\}, \quad k = 2,\dots,C, \end{aligned}$$

and univariate Lipschitz continuous functions $\rho_k : \mathbb{R} \to \mathbb{R}$, $k = 2,\dots,C$, and $\varphi_k : \mathbb{R} \to \mathbb{R}$, $k = 1,\dots,C$. Alternatively, a recursive formulation of (4) is

$$f = f^C, \quad f^1 = \varphi_1 \circ f_1, \quad f^k = \rho_k \circ g_k(f^{k-1}, \varphi_k \circ f_k),\ k \geq 2.$$

This allows for a rather large class of constraint functions. For instance, (1) is recovered for $g_k = \max$, $k = 2,\dots,C$, and the boolean expression

$$(f_1(x,\theta) \leq 0\ \vee\ f_2(x,\theta) \leq 0)\ \wedge\ f_3(x,\theta) \leq 0$$

is equivalent to $\max\{\min\{f_1(x,\theta), f_2(x,\theta)\}, f_3(x,\theta)\} \leq 0$. A polynomial $f_k(x,\theta)$ can also be considered with $\phi_k(x)$ gathering all its monomials and $\psi_k(\theta)$ its coefficients.

We note that, while similar in some aspects, our framework is more general than the structure imposed on the constraints by [12], which is basically limited to the form (1)–(2) with the maximum operator and constant dimensions $n_1 = \dots = n_C$, a *convex* set $\mathcal{X}$ and *convex* functions ϕ_k.

A. Standard scenario approach

The standard scenario approach aims at approximating the solution to Problem 1 by solving the following version with a finite number of sampled constraints.

Problem 2 (Scenario program)**.** *Given a probability distribution of $\theta \in \Theta$, draw a random sample of N independent parameter values $(\theta_i)_{1\leq i\leq N} \subset \Theta$ and solve*

$$\hat{x} = \underset{x\in\mathcal{X}}{\operatorname{argmin}}\, J(x), \quad s.t.\ f(x,\theta_i) \leq 0, \quad i = 1,\dots,N.$$

The probability distribution of θ can be given by a priori knowledge, the uncertainties in the modeling process, or is typically taken as the uniform distribution over Θ which is often the "distributionally robust" choice in the sense of [14].

The ability of Problem 2 to provide solutions that satisfy the infinite number of constraints of Problem 1 is assessed via the risk, i.e., the probability of violation

$$V(x) = P\{f(x,\theta) > 0\}, \tag{5}$$

where P stands for the probability distribution of θ. We also define the empirical risk as the fraction of violated constraints among the scenarios, i.e.,

$$\hat{V}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{f(x,\theta_i)>0}. \tag{6}$$

A classical result [2], [4] guarantees that for *convex* problems with $\mathcal{X} \subset \mathbb{R}^d$ and $\hat{x}$ the solution to Problem 2,

$$P^N\{V(\hat{x}) \leq \epsilon\} \geq 1 - \delta, \tag{7}$$

where P^N stands for the product probability distribution of $(\theta_i)_{1\leq i\leq N}$ and

$$\delta = \sum_{j=0}^{d-1} \binom{N}{j} \epsilon^j (1-\epsilon)^{N-j}. \tag{8}$$

Thus, with high probability on the random draw of the scenarios $(\theta_i)_{1\leq i\leq N}$, the risk can be guaranteed to be as small as ϵ. Such results can be used to compute the sample complexity, i.e., the minimal number of scenarios N required to guarantee a priori a certain performance for given ϵ and δ.

B. Introducing the margin

We first focus on feasibility problems, i.e., those for which no J is to be minimized, and the guarantees regarding the probability of violation.

In this paper, we introduce the notion of *margin* and the following hard-margin version of Problem 2.

Problem 3 (Hard-margin scenario program)**.** *Given a probability distribution of $\theta \in \Theta$ and a margin parameter $\gamma > 0$, draw a sample of N independent parameter values $(\theta_i)_{1\leq i\leq N} \subset \Theta$ and*

$$\textit{Find } x \in \mathcal{X}, \quad s.t.\ f(x,\theta_i) \leq -\gamma, \quad i = 1,\dots,N.$$

In Problem 3, the constraints are enforced with an additional margin that will ease the satisfaction of the original constraint $f(x,\theta) \leq 0$ when generalizing to other values of θ, as illustrated by Fig. 1.

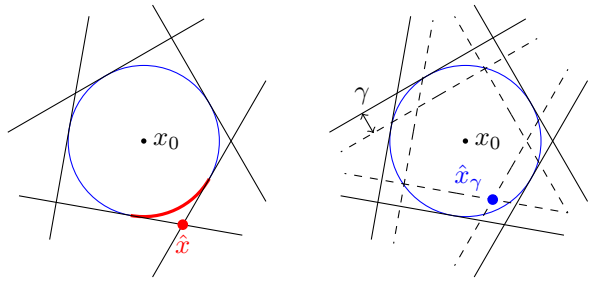

Fig. 1. Illustration of the benefit of the margin on a toy example: the constraint $\|x - x_0\| \leq 1$ (blue circle) is implemented by an infinite number of linear constraints, $f(x,\theta) = \theta^\top(x - x_0) - 1 \leq 0$, for θ uniformly distributed in the unit circle. *Left:* The standard scenario approach yields a solution $\hat{x}$ that satisfies the 5 scenario constraints (the 5 lines) but has a probability of violation of $V(\hat{x}) = 20\%$ that corresponds to the probability of drawing a tangent line to the circle at a point along its red part. *Right:* All feasible solutions to the hard-margin scenario Problem 3, such as $\hat{x}_\gamma$, satisfy the 5 scenario constraints with a margin γ (the 5 dashed lines) and also satisfy the worst-case constraint (blue circle) with $V(\hat{x}_\gamma) = 0$.

When Problem 3 (or more generally Problem 1) is infeasible, slack variables can be introduced to allow for small violations of the margin constraints, leading to a soft-margin version of the approach. This is reminiscent of the relaxation approach suggested in [8] for allowing the violation of some scenario constraints.

Problem 4 (Soft-margin scenario program)**.** *Given a probability distribution of $\theta \in \Theta$ and a margin parameter $\gamma > 0$, draw a sample of N independent parameter values $(\theta_i)_{1\leq i\leq N} \subset \Theta$ and solve*

$$\min_{x\in\mathcal{X},\xi\in\mathbb{R}^N} \sum_{i=1}^{N} \xi_i \quad s.t. \begin{cases} f(x,\theta_i) \leq -\gamma + \xi_i, & i = 1,\dots,N \\ \xi_i \geq 0, & i = 1,\dots,N. \end{cases}$$

As illustrated by Fig. 2, the soft margin can be larger than the hard margin, which will translate into better performance guarantees in the analysis below.

Note that, in the case of (1), the scales of the f_k's are meaningless for $f(x,\theta) \leq 0$ but they impact the solution of Problems 3–4. In such cases, the f_k's should be normalized.

III. Analysis

In order to refine the analysis of the quality of the solution of Problems 3–4, we introduce margin counterparts to the risks. These are based on a margin loss function $\ell_\gamma : \mathcal{X}\times\Theta \to [0,1]$ which measures the ability of x to satisfy the margin constraint $f(x,\theta) \leq -\gamma$ for a single value of θ. More precisely, we consider the piecewise linear margin loss function given by

$$\ell_\gamma(x,\theta) = \begin{cases} 1, & \text{if } f(x,\theta) \geq 0 \\ 1 + \frac{f(x,\theta)}{\gamma}, & \text{if } f(x,\theta) \in (-\gamma, 0) \\ 0, & \text{if } f(x,\theta) \leq -\gamma \end{cases} \tag{9}$$

and define the risk at margin γ as

$$V_\gamma(x) = \mathbb{E}_\theta \ell_\gamma(x,\theta),$$

and the empirical margin risk as $\hat{V}_\gamma(x) = \frac{1}{N}\sum_{i=1}^{N} \ell_\gamma(x,\theta_i)$.

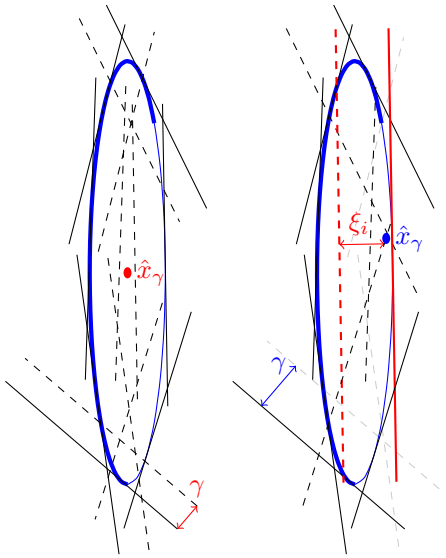

Fig. 2. Soft-margin approach illustrated on an example where random linear constraints are substituted for a nonlinear constraint (blue ellipsoid). *Left:* The hard-margin scenario program yields a solution $\hat{x}_\gamma$ that is feasible for all $\theta \in \Theta$ with $V(\hat{x}_\gamma) = 0$. However, the margin γ is rather small and cannot be increased without making Problem 3 infeasible. Thus, the upper bound on $V(\hat{x}_\gamma)$ given by Theorem 1 is not really tight and, though the solution is perfect, we have limited confidence in its performance. *Right:* By allowing the violation of one margin constraint (the red dashed line) with a slack of ξ_i, the soft-margin version yields a solution with a larger margin γ, and thus with more confidence in its performance despite the increase in the empirical error $\hat{V}_\gamma(\hat{x}_\gamma)$.

In comparison with the standard definition of the risk, the margin risk $V_\gamma(x)$ based on the loss (9) also accounts for scenarios on which the constraint is satisfied with not enough margin. It is easy to verify that

$$\forall x \in \mathcal{X}, \theta \in \Theta, \quad \ell_\gamma(x,\theta) \geq \mathbf{1}_{f(x,\theta)>0} \tag{10}$$

and therefore that $V_\gamma(x) \geq \mathbb{E}_\theta \mathbf{1}_{f(x,\theta)>0} = V(x)$ always holds. Thus, by upper bounding the margin risk $V_\gamma(x)$, we will also bound the true probability of violation $V(x)$. Though this may introduce some slack in the bound, this allows for a finer analysis since $V_\gamma(x)$ depends more precisely on the values of $f(x,\theta)$ and not just on its sign as $V(x)$.

A. Bounds on the probability of violation

Inspired by the statistical learning approach initiated by [10], [11], the behavior of the margin risk is here analyzed in terms of the Rademacher complexity.

Definition 1 (Rademacher complexity)**.** *For $N \in \mathbb{N}^*$, let $\boldsymbol{\theta}_N = (\theta_i)_{1\leq i\leq N}$ be an N-sample of independent copies of the random variable $\theta \in \Theta$, let $\boldsymbol{\sigma}_N = (\sigma_i)_{1\leq i\leq N}$ be a sequence of independent random variables uniformly distributed in $\{-1,+1\}$. Let $\mathcal{F}$ be a class of real-valued functions on Θ. The* empirical Rademacher complexity *of $\mathcal{F}$ given $\boldsymbol{\theta}_N$ is*

$$\hat{\mathcal{R}}_N(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_N}\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \sigma_i f(\theta_i) \,\middle|\, \boldsymbol{\theta}_N\right].$$

The Rademacher complexity *of $\mathcal{F}$ is $\mathcal{R}_N(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\theta}_N}\hat{\mathcal{R}}_N(\mathcal{F})$.*

In particular, we have the following generic theorem, that applies similarly to the solution of Problem 2, 3 or 4.

Theorem 1. *For any $\gamma > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$ on the random draw of $(\theta_i)_{1\leq i\leq N}$, the probability of violation is uniformly bounded for all $x \in \mathcal{X}$ by*

$$V(x) \leq V_\gamma(x) \leq \hat{V}_\gamma(x) + \frac{2}{\gamma}\mathcal{R}_N(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2N}},$$

where

$$\mathcal{F} = \{f_x \in \mathbb{R}^\Theta : f_x(\theta) = f(x,\theta),\ x \in \mathcal{X}\}. \tag{11}$$

Proof. As already pointed out, the first inequality is a direct consequence of (10). The second inequality stems from Theorem 3.3 in [15] applied to the class of loss functions

$$\mathcal{L}_\gamma = \{\ell_{\gamma,x} \in [0,1]^\Theta : \ell_{\gamma,x}(\theta) = \ell_\gamma(x,\theta),\ x \in \mathcal{X}\}, \tag{12}$$

and which guarantees that, with probability at least $1-\delta$,

$$V_\gamma(x) \leq \hat{V}_\gamma(x) + 2\mathcal{R}_N(\mathcal{L}_\gamma) + \sqrt{\frac{\log\frac{1}{\delta}}{2N}}. \tag{13}$$

Then, the Rademacher complexity of the loss class (12) is bounded in terms of the class $\mathcal{F}$ using the contraction principle (see Lemma 5.7 in [15]) together with the fact that ℓ_γ can be written as $\ell_\gamma(x,\theta) = \varphi \circ f(x,\theta)$ with the function $\varphi(t) = \min\{1, \max\{0, 1+t/\gamma\}\}$ of Lipschitz constant equal to $1/\gamma$. This leads to $\mathcal{R}_N(\mathcal{L}_\gamma) \leq \mathcal{R}_N(\mathcal{F})/\gamma$ and concludes the proof. □

Theorem 1 allows one to bound the risk in terms of the empirical risk with a confidence interval that depends on the complexity of the constraint functions induced by the domain $\mathcal{X}$. This is very much related to the approach of [3], which used the Vapnik-Chervonenkis (VC) dimension of $\mathcal{L} = \{\ell_x \in [0,1]^\Theta : \ell_x(\theta) = \mathbf{1}_{f(x,\theta)>0},\ x \in \mathcal{X}\}$ to measure this complexity. Here, the Rademacher complexity offers a finer measure of capacity that also takes into account the magnitude of $f(x,\theta)$, whereas the VC-dimension of $\mathcal{L}$ only depends on its sign. However, the two can be related (see, e.g., [15]): for a class $\mathcal{L}$ of finite VC-dimension d_{VC}, results in the flavor of those of [3] can be recovered in our general approach by substituting $\mathcal{L}$ for the loss class $\mathcal{L}_\gamma$ in the proof above and bounding its Rademacher complexity by

$$\mathcal{R}_N(\mathcal{L}) \leq \sqrt{\frac{2d_{VC}\log\frac{eN}{d_{VC}}}{N}}. \tag{14}$$

An important feature of Theorem 1 is that it holds uniformly over $\mathcal{X}$, meaning that it applies whether one truly solves Problems 3–4 or not, as long as the computed solution $\hat{x}_\gamma$ lies in the domain $\mathcal{X}$. This is crucial for nonconvex instances for which true solutions are difficult to obtain, and stands in contrast to the standard results discussed in the literature that only apply to the optimizer of the scenario problem [4], [5] or to solutions that at least satisfy all constraints [7].

The influence of the margin γ on the performance guarantees is clear from Theorem 1: the confidence interval of the upper bound on the probability of violation decreases linearly

with $1/\gamma$. Thus, a larger margin ensures better confidence, as illustrated by Fig. 2. In the right plot of this Figure, the empirical error $\hat{V}_\gamma(\hat{x}_\gamma)$ increases by $\xi_i/\gamma N$ as one of the margin constraints is violated.[1] Meanwhile, the increase of the margin decreases the confidence interval and, overall, leads to a more favorable risk bound than in the left plot. Here, the thick part of the ellipsoid represents the points where tangent lines are drawn with high probability, while the thin part corresponds to random constraints that occur with low probability. Thus, the red constraint is a rare event and violating it does not incur a large increase of the probability of violation, which explains why a solution with more errors can lead to better guarantees in generalization. The low probability also means that there are fewer such constraints in the scenarios and that the cost of Problem 4 is minimized when violating the red constraint rather than multiple other constraints.

In order to produce a performance guarantee, the Rademacher complexity appearing in Theorem 1 must be estimated. The following theorem (proved in Appendix) shows how to upper bound this complexity for constraint functions as in (4).

Theorem 2. *Let $\mathcal{F}$ be a function class as in* (11) *with f of the form* (4)*. Then,*

$$\mathcal{R}_N(\mathcal{F}) \leq \sum_{k=1}^{C} \frac{(\prod_{j=k}^{C} \overline{\rho}_j)\overline{\varphi}_k \tau_k \Lambda_k}{\sqrt{N}},$$

where $\tau_k = \sup_{\theta\in\Theta} \|\psi_k(\theta)\|$, $\Lambda_k = \sup_{x\in\mathcal{X}} \|\phi_k(x)\|$, $\overline{\rho}_1 = 1$, $\overline{\rho}_k$ for $k \geq 2$ and $\overline{\varphi}_k$ for $k \geq 1$ are the Lipschitz constants of the functions ρ_k and φ_k, respectively.

Theorem 2 shows that the complexity, and thus the generalization performance via Theorem 1, does not depend on the dimension of x but rather on τ_k, which measures the size of the $\psi_k(\theta)$'s over the uncertainty set Θ, and Λ_k, which measures the size of the $\phi_k(x)$'s over $\mathcal{X}$. This stands in contrast to most of the literature on scenario optimization, where the dimension plays a major role. This can be seen in (8) for convex problems as studied in [4], while for the approach of [3], d_{VC} in (14) grows linearly with d even for a simple constraint function (4) with $C = 1$, φ_1 the identity and f_1 a linear function of x. When considering integer programming, the method of [5] also leads to a bound linear in the number of variables. Finally, the approach of [12] for structured convex constraints yields bounds that do not depend on d but explicitly involve the dimension n_k of $\phi_k(x)$ instead.

The parameters τ_k in Theorem 2 reflect the size of the uncertainties, which in contrast are ignored by the standard approaches that yield the same level of guarantee for small and large uncertainties. The parameters Λ_k measure the size of the $\phi_k(x)$'s, which are related to the size of the domain $\mathcal{X}$, and thus they encode our prior knowledge on the solution. In particular, what really matters is the radius of the smallest ball enclosing the $\phi_k(x)$'s induced by all $x \in \mathcal{X}$, as made precise in the following corollary. Therefore, if the search domain $\mathcal{X}$ can be a priori reduced in size (for instance after the computation of a preliminary good guess of the solution), the radius could typically be made smaller, leading to better guarantees on the probability of violation.

Corollary 1. *Let $\mathcal{F}$ be a function class as in* (11) *with f as in* (4)*. Then, for any set of centers $(\phi_{0,k})_{1\leq k\leq C} \in \prod_{k=1}^{C} \mathbb{R}^{n_k}$,*

$$\mathcal{R}_N(\mathcal{F}) \leq \sum_{k=1}^{C} \frac{(\prod_{j=k}^{C} \overline{\rho}_j)\overline{\varphi}_k \tau_k \tilde{\Lambda}_k}{\sqrt{N}},$$

where $\tilde{\Lambda}_k = \sup_{x\in\mathcal{X}} \|\phi_k(x) - \phi_{0,k}\|$ and the other constants are as in Theorem 2.

Proof. Given a choice of $(\phi_{0,k})_{1\leq k\leq C}$, reformulate the f_k's as

$$\begin{aligned} f_k(x,\theta) &= \psi_k(\theta)^\top(\phi_k(x) - \phi_{0,k}) + (\eta_k(\theta) + \psi_k(\theta)^\top \phi_{0,k}) \\ &= \psi_k(\theta)^\top \tilde{\phi}_k(x) + \tilde{\eta}_k(\theta) \end{aligned}$$

and note that in the proofs of Theorem 2 and Lemma 1, the definition of $\eta_k(\theta)$ does not play any role (as long as it does not depend on x). Therefore, they hold similarly with $\tilde{\phi}_k(x)$, $\tilde{\eta}_k(\theta)$ instead of $\phi_k(x)$, $\eta_k(\theta)$, and $\sup_{x\in\mathcal{X}} \|\phi_k(x)\|$ replaced by $\sup_{x\in\mathcal{X}} \|\tilde{\phi}_k(x)\| = \sup_{x\in\mathcal{X}} \|\phi_k(x) - \phi_{0,k}\| = \tilde{\Lambda}_k$. □

The combination of Theorems 1 and 2 (or Corollary 1) leads to a bound on the probability of violation that converges to zero as $O(1/\sqrt{N})$, which might appear weaker than most results in the literature on scenario optimization, such as (7)–(8), which typically lead to bounds in $O(d/N)$. However, a bound in $O(d/N)$ that improves upon Theorems 1–2 requires that $d = O(\sqrt{N})$, which fails in high-dimensional regimes. This informal argument in favor of the proposed approach for high-dimensional problems, and even convex ones, will be made precise below when discussing the sample complexity.

Regarding the linear dependence of the bounds with respect to C, it is reminiscent of the fact that the constraint function (4) is built from a combination of several constraints/components. A similar effect can be observed in the results of [12] for convex problems based on (1).[2]

Finally, recall that Theorem 1 holds uniformly over $\mathcal{X}$, which is particularly desirable for nonconvex problems to bypass the assumption that the computed solution is optimal (or even deterministic). However, this introduces the constants τ_k and Λ_k in Theorem 2 and can prevent us from obtaining satisfactory guarantees when their values are too large, e.g., when one lacks prior knowledge on the location of the solution or when the uncertainties can have a large magnitude. In such cases, standard approaches like [2], [4] might be preferrable.

B. Sample complexity

The sample complexity of a scenario problem is the minimal number of scenarios that are required to guarantee that, with high probability, the probability of violation does not exceed a predefined threshold. The following sample complexity estimate can be derived from the results in Theorems 1 and 2.

[1] Note that Problem 4 incurs a linear cost for errors ξ_i, whereas the loss function (9) saturates at 1. Thus, in general, each error incurs a loss $\min(\xi_i/\gamma, 1)/N$ in $\hat{V}_\gamma(x)$, while in Fig. 2, $\xi_i < \gamma$.

[2] The term $\prod_{j=k}^{C} \overline{\rho}_j$ is not considered here as it typically is a product of values close to one, and is even exactly 1 for all problems analyzed in [12].

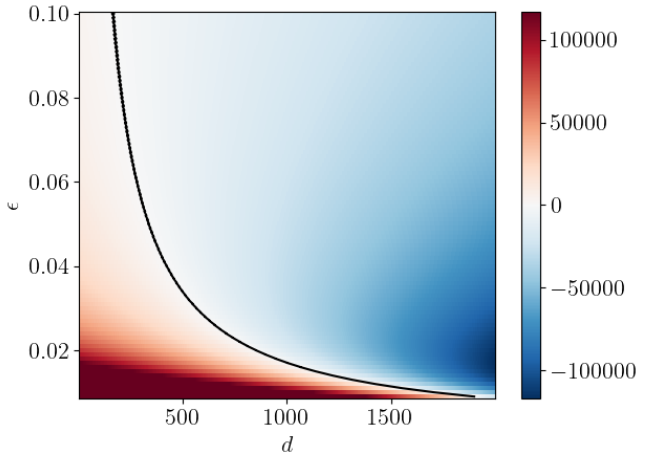

Fig. 3. Color plot of $N(\epsilon,\delta) - N_{\text{cvx}}(\epsilon,\delta)$ as a function of d and ϵ for $\delta = 0.001$, $C = 1$, $\overline{\rho}_1\overline{\varphi}_1\tau_1\Lambda_1/\gamma = 2$. The black line shows the boundary between the region where $N(\epsilon,\delta) < N_{\text{cvx}}(\epsilon,\delta)$ (in blue) and the one where $N(\epsilon,\delta) > N_{\text{cvx}}(\epsilon,\delta)$ (in red).

Corollary 2. *Let f be of the form* (4)*. Then, for any N larger than or equal to the sample complexity*

$$N(\epsilon,\delta) = \frac{\left(\frac{2}{\gamma}\sum_{k=1}^{C}(\prod_{j=k}^{C}\overline{\rho}_j)\overline{\varphi}_k\tau_k\Lambda_k + \sqrt{\log\frac{1}{\delta^{1/2}}}\right)^2}{\epsilon^2}, \tag{15}$$

the probability of violation $V(x)$ of any $x \in \mathcal{X}$ is less than or equal to $\hat{V}_\gamma(x) + \epsilon$ with probability at least $1-\delta$. In addition, if Problem 3 is feasible, then all its solutions $\hat{x}_\gamma$ satisfy $\hat{V}_\gamma(\hat{x}_\gamma) = 0$ and the bound $V(\hat{x}_\gamma) \leq \epsilon$ holds with probability at least $1-\delta$ for any $N \geq N(\epsilon,\delta)$.

Though the sample complexity thus obtained is quadratic wrt. $1/\epsilon$, its numerical value can be less than that obtained with the standard approach for convex problems in high dimension. Indeed, as shown in [4], for convex problems, (7)–(8) translate into the sample complexity

$$N_{\text{cvx}}(\epsilon,\delta) = \frac{2d + 2\log\frac{1}{\delta}}{\epsilon}, \tag{16}$$

which grows linearly with the dimension d. Therefore, there is a d above which $N(\epsilon,\delta) < N_{\text{cvx}}(\epsilon,\delta)$. Figure 3 shows the region where the proposed approach thus improves upon [4] in the (d,ϵ) space for a fixed value of the bound in Theorem 2.

C. Maximizing the margin

An insight from Theorem 1 is that better generalization results from a larger margin γ. Therefore, it would be tempting to replace Problem 3 by:

$$\max_{x\in\mathcal{X},\gamma>0} \gamma, \quad \text{s.t. } f(x,\theta_i) \leq -\gamma, \quad i = 1,\ldots,N. \tag{17}$$

However, Theorem 1 only holds for a value of γ fixed a priori and cannot be applied with the margin resulting from (17) and that depends on the random data $(\theta_i)_{1\leq i\leq N}$. Nonetheless, it can be extended to hold uniformly over γ by following the proof of Theorem 5.9 in [15], with the addition of a small term in $O(\sqrt{\log\log_2(1/\gamma)})$. Therefore, (17) could be used to estimate a value of γ that produces a better bound on the probability of violation.

D. A posteriori bounds

In cases where the distribution of θ is unknown or samples cannot be easily generated, the standard Monte Carlo approach to assess the probability of violation on an independent and large sample is not available. Then, a posteriori bounds can be used to estimate the probability of violation $V(x)$ of a solution x computed with a fixed budget of N observations $(\theta_i)_{1\leq i\leq N}$ by the empirical error $\hat{V}(x)$ computed on the same data.

While the bounds on $V(x)$ in the theorems above could be applied for this purpose, they rely on a worst-case strategy in order to be computable a priori or to yield sample complexity estimates. Instead, refined a posteriori bounds can be computed from the *observed value* of the sample $(\theta_i)_{1\leq i\leq N}$ using the *empirical* Rademacher complexity. In this case, Theorem 1 holds with only minor changes in the constants, and Theorem 2 can be reshaped with $\sqrt{\sum_{i=1}^{N}\|\psi_k(\theta_i)\|^2}/N$ substituted for $\tau_k/\sqrt{N}$. In addition, the constants Λ_k can also be replaced by empirical versions computed specifically for the solution $\hat{x}_\gamma$ returned by the algorithm rather than global upper bounds.

E. Fast rates

As already mentioned, most of the literature on scenario optimization concentrates on bounds on the probability of violation that are in $O(1/N)$ rather than $O(1/\sqrt{N})$. Rademacher complexity-based bounds as proposed here can be extended using localization arguments [16], [17] to yield bounds in $O(1/N)$ that are still independent of the dimension d of x. Alternatively, covering numbers could be used instead of the Rademacher complexity to lead to bounds in $O(\log(N)/N)$, by following [18] and using results of [19].

IV. Optimization problems and margin complexity

First of all, notice that all results presented above hold similarly for random optimization problems of the form:

$$\min_{x\in\mathcal{X}} J(x), \quad \text{s.t. } f(x,\theta_i) \leq -\gamma, \quad i = 1,\ldots,N. \tag{18}$$

However, introducing the margin can affect the objective value $J(\hat{x}_\gamma)$ of the computed solution $\hat{x}_\gamma$. Since the maximization of the margin as proposed in Sect. III-C and the minimization of $J(x)$ play against each other, one would have to find a suitable trade-off.

Another point of view arises when one considers the following question: given a fixed budget of N scenarios, how can we minimize $J(x)$ while ensuring that $P^N\{V(x) \leq \epsilon\} \geq 1-\delta$? Clearly, the smallest $J(\hat{x}_\gamma)$ will result from the smallest choice of margin γ in (18). This leads to the concept of *margin complexity*, i.e., the smallest margin γ such that $P^N\{V(x) \leq \epsilon\} \geq 1-\delta$ for given N, ϵ and δ. For constraint functions as in (4) and the hard-margin scenario program (18) that yields

solutions such that $\hat{V}_\gamma(\hat{x}_\gamma) = 0$, the margin complexity can be estimated using Theorems 1–2 as

$$\gamma(N, \epsilon, \delta) = \frac{2\sum_{k=1}^{C}(\prod_{j=k}^{C}\overline{\rho}_j)\overline{\varphi}_k\tau_k\Lambda_k}{\epsilon\sqrt{N} - \sqrt{\log\frac{1}{\delta^{1/2}}}}.$$

The procedure to handle an optimization problem with a fixed budget can thus be formulated as: solve (18) with $\gamma = \gamma(N, \epsilon, \delta)$ given above. Then, we have the guarantee that the solution $\hat{x}_\gamma$ satisfies $P^N\{V(\hat{x}_\gamma) \le \epsilon\} \ge 1 - \delta$.

V. Numerical example

Consider a convex problem with cost function $J(x) = c^\top x$ with all entries of c set to -1, the domain $\mathcal{X} \subset \mathbb{R}^d$ defined as the unit ball of center $(1 + 1/\sqrt{d})x_0$, uncertainties θ sampled from the section of the unit sphere in the nonnegative orthant of $\mathbb{R}^d$ and $f(x, \theta) = \theta^\top(x - x_0) - 1$ with x_0 a vector of all ones. We compare the results of the approach of [2], that computes $\hat{x}$, and the proposed method (18), that uses the margin γ to compute $\hat{x}_\gamma$, over various numbers N of scenarios with $d = 500$ and $\delta = 0.01$. Table I reports ϵ and ϵ_γ, which are the upper bounds on $V(\hat{x})$ and $V(\hat{x}_\gamma)$, and $\epsilon^*(\cdot)$, which is the Monte Carlo estimate of $V(\cdot)$ over an independent sample of 100000 values of θ (a $\star$ means that the solution truly satisfies the robust constraint $\forall\theta \in \Theta$, whereas 0.000 indicates that $V(\cdot)$ is very small).[3] We also report the relative optimality gap, $|(J(\hat{x}_\gamma) - J(\hat{x}))/J(\hat{x})|$, to measure the increase of the cost induced by the use of the margin γ; and show how it improves when setting $\gamma = \gamma(N, \epsilon, \delta)$ as proposed in Sect. IV for the value of ϵ given by [2].

These results show that on this high-dimensional problem the proposed method yields both better guarantees and smaller actual risks than the standard approach of [2]: $\epsilon_\gamma < \epsilon$ and $\epsilon^*(\hat{x}_\gamma) < \epsilon^*(\hat{x})$, though ϵ of [2] has a smaller relative error in estimating $\epsilon^*(\hat{x})$ than the proposed ϵ_γ for $\epsilon^*(\hat{x}_\gamma)$. Also, the cost of the solution increases only by a small amount (though the significance of this increase may depend on the application). In addition, when switching to a nonconvex version of the problem over a quantized domain $\mathcal{X}' \subset \mathcal{X}$ with each entry of x taking one of 10 discrete values, the bounds ϵ_γ of the proposed method hold similarly, whereas the extension of [2] proposed in [5] has to be considered and leads to the increased values of ϵ reported in the last row of Table I.

TABLE I
Numerical comparison of the proposed method with the standard ones of [2] and [5]. See Sect. V for details.

Method	N	1000	3000	10000	20000
[2]	ϵ	> 1	0.336	0.101	0.050
	$\epsilon^*(\hat{x})$	0.447	0.154	0.047	0.024
This paper	ϵ_γ	0.174	0.101	0.082	0.039
with	$\epsilon^*(\hat{x}_\gamma)$	0.000	$\star$	$\star$	$\star$
$\gamma = 0.5$	Optimality gap	2.64%	2.61%	1.55%	2.58%
This paper	γ	0.06	0.12	0.23	0.36
with	$\epsilon^*(\hat{x}_\gamma)$	0.006	0.000	$\star$	$\star$
$\gamma(N, \epsilon, \delta)$	Optimality gap	0.35%	0.62%	1.21%	1.84%
[5] with $\mathcal{X}'$	ϵ	> 1	0.771	0.231	0.115

VI. Conclusions

The paper introduced the notion of margin in the scenario approach to robust optimization and analyzed the statistical performance of the resulting random programs using tools from statistical learning theory, and more specifically the Rademacher complexity. This allowed for the derivation of a priori bounds on the probability of violation and sample complexities in which the problem dimension could be removed. Instead, these quantities depend on how parameters related to the magnitude of the uncertainties and the size of the search domain compare with the level of margin with which the scenario constraints are satisfied. On the one hand, these results hold only for a specific class of constraint functions that includes polynomials as particular cases, but on the other hand they apply similarly to both convex and nonconvex problems.

Future work will consider further generalizations of the class of problems for which such results can be derived. While the main focus of the paper was on feasibility problems and the probability of violation, the impact of the margin on the cost function for minimization problems should also be further investigated.

[3] Given 100000 trials without errors, Bernstein inequality yields that $V(\cdot)$ is less than 0.0002 with probability at least 99%.

References

[1] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized algorithms for analysis and control of uncertain systems: with applications*. Springer, 2013.

[2] 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.

[3] 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.

[4] G. Calafiore, "Random convex programs," *SIAM Journal on Optimization*, vol. 20, no. 6, pp. 3427–3464, 2010.

[5] P. Esfahani, T. Sutter, and J. Lygeros, "Performance bounds for the scenario approach and an extension to a class of non-convex programs," *IEEE Transactions on Automatic Control*, vol. 60, no. 1, pp. 46–58, 2014.

[6] M. Campi and S. Garatti, "Wait-and-judge scenario optimization," *Mathematical Programming*, vol. 167, no. 1, pp. 155–189, 2018.

[7] M. Campi, S. Garatti, and F. Ramponi, "A general scenario theory for nonconvex optimization and decision making," *IEEE Transactions on Automatic Control*, vol. 63, no. 12, pp. 4067–4078, 2018.

[8] S. Garatti and M. Campi, "Risk and complexity in scenario optimization," *Mathematical Programming*, vol. 191, pp. 243–279, 2022.

[9] 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.

[10] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," *The Annals of Statistics*, vol. 30, no. 1, pp. 1–50, 2002.

[11] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[12] X. Zhang, S. Grammatico, G. Schildbach, P. Goulart, and J. Lygeros, "On the sample size of random convex programs with structured dependence on the uncertainty," *Automatica*, vol. 60, pp. 182–188, 2015.

[13] S. Garatti, A. Carè, and M. Campi, "Complexity is an effective observable to tune early stopping in scenario optimization," *IEEE Transactions on Automatic Control*, vol. 68, no. 2, pp. 928–942, 2023.

[14] B. Barmish and C. Lagoa, "The uniform distribution: A rigorous justification for its use in robustness analysis," *Mathematics of Control, Signals, and Systems*, vol. 10, pp. 203–222, 1997.

[15] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. The MIT Press, 2018.

[16] P. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *The Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[17] N. Srebro, K. Sridharan, and A. Tewari, "Smoothness, low noise and fast rates," in *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[18] C. Cortes, M. Mohri, and A. Suresh, "Relative deviation margin bounds," in *ICML*, 2021, pp. 2122–2131.

[19] T. Zhang, "Covering number bounds of certain regularized linear function classes," *Journal of Machine Learning Research*, vol. 2, pp. 527–550, 2002.

APPENDIX

The proof of Theorem 2 relies on the following bound on the Rademacher complexity of a class of functions f_k as in (2).

Lemma 1. *Given functions $\psi : \Theta \to \mathbb{R}^n$, $\phi : \mathcal{X} \to \mathbb{R}^n$ and $\eta : \Theta \to \mathbb{R}$, the empirical Rademacher complexity of the class $\mathcal{F} = \{f_x \in \mathbb{R}^\Theta : f_x(\theta) = \psi(\theta)^\top \phi(x) + \eta(\theta),\ x \in \mathcal{X}\}$ given $(\theta_i)_{1 \leq i \leq N}$ is bounded by*

$$\hat{\mathcal{R}}_N(\mathcal{F}) \leq \frac{\Lambda \sqrt{\sum_{i=1}^N \|\psi(\theta_i)\|^2}}{N} \leq \frac{\tau \Lambda}{\sqrt{N}},$$

where $\tau = \sup_{\theta \in \Theta} \|\psi(\theta)\|$ and $\Lambda = \sup_{x \in \mathcal{X}} \|\phi(x)\|$.

Proof. Using the subadditivity of the supremum, we have

$$\begin{aligned}\hat{\mathcal{R}}_N(\mathcal{F}) &= \mathbb{E} \sup_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N \sigma_i (\psi(\theta_i)^\top \phi(x) + \eta(\theta_i)) \\ &\leq \mathbb{E} \sup_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N \sigma_i \psi(\theta_i)^\top \phi(x) + \mathbb{E} \sup_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N \sigma_i \eta(\theta_i),\end{aligned}$$

where $\mathbb{E} \sup_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N \sigma_i \eta(\theta_i) = \mathbb{E} \frac{1}{N} \sum_{i=1}^N \sigma_i \eta(\theta_i) = \frac{1}{N} \sum_{i=1}^N \eta(\theta_i) \mathbb{E} \sigma_i = 0$, since the expectation is computed only with respect to the Rademacher variables σ_i that are centered ($\mathbb{E}\sigma_i = 0$). We can then follow the standard path for linear classes [11], with the addition of the mappings ψ and ϕ:

$$\begin{aligned}\hat{\mathcal{R}}_N(\mathcal{F}) &\leq \mathbb{E} \sup_{x \in \mathcal{X}} \frac{1}{N} \sum_{i=1}^N \sigma_i \psi(\theta_i)^\top \phi(x) \\ &= \frac{1}{N} \mathbb{E} \sup_{x \in \mathcal{X}} \left(\sum_{i=1}^N \sigma_i \psi(\theta_i) \right)^\top \phi(x) \\ &\leq \frac{1}{N} \mathbb{E} \sup_{x \in \mathcal{X}} \left\| \sum_{i=1}^N \sigma_i \psi(\theta_i) \right\| \|\phi(x)\| \\ &= \frac{\Lambda}{N} \mathbb{E} \left\| \sum_{i=1}^N \sigma_i \psi(\theta_i) \right\|,\end{aligned}$$

in which Jensen's inequality ensures that

$$\begin{aligned}\mathbb{E} \left\| \sum_{i=1}^N \sigma_i \psi(\theta_i) \right\| &\leq \sqrt{\mathbb{E} \left\| \sum_{i=1}^N \sigma_i \psi(\theta_i) \right\|^2} \\ &= \sqrt{\mathbb{E} \sum_{i=1}^N \sum_{j=1}^N \sigma_i \sigma_j \psi(\theta_i)^\top \psi(\theta_j)} \\ &= \sqrt{\sum_{i=1}^N \|\psi(\theta_i)\|^2 + \mathbb{E} \sum_{j \neq i} \sigma_i \sigma_j \psi(\theta_i)^\top \psi(\theta_j)} \\ &= \sqrt{\sum_{i=1}^N \|\psi(\theta_i)\|^2},\end{aligned}$$

where we used the independence of the σ_i's and the fact that $\mathbb{E}\sigma_i = 0$ to remove the terms with $j \neq i$. Gathering the results and using $\sqrt{\sum_{i=1}^N \|\psi(\theta_i)\|^2} \leq \sqrt{N}\tau$ complete the proof. □

The proof of Theorem 2 also requires the introduction of a few function classes: for all $k \in \{1, \dots, C\}$,

$$\mathcal{F}_k = \{f_x \in \mathbb{R}^\Theta : f_x(\theta) = f_k(x, \theta),\ x \in \mathcal{X}\}$$

and, for all $k \in \{2, \dots, C\}$,

$$\mathcal{G}_k(\mathcal{U}, \mathcal{V}) = \mathcal{U} + \mathcal{V},\ \mathcal{U} - \mathcal{V},\ \max(\mathcal{U}, \mathcal{V}),\ \text{or}\ \min(\mathcal{U}, \mathcal{V}),$$

in accordance with g_k. With these classes at hand, we can rewrite the function class of interest, $\mathcal{F}$, in a recursive manner:

$$\begin{aligned}\mathcal{F} = \mathcal{F}^C, \quad \mathcal{F}^1 = \varphi_1 \circ \mathcal{F}_1 \\ \mathcal{F}^{k+1} = \rho_{k+1} \circ \mathcal{G}_{k+1}(\mathcal{F}^k,\ \varphi_{k+1} \circ \mathcal{F}_{k+1}).\end{aligned}$$

Then, the proof works by induction over the number of components C. First, for $C = 1$, $f(x, \theta) = \varphi_1 \circ f_1(x, \theta)$ and $\mathcal{F} = \varphi_1 \circ \mathcal{F}_1$. In this case, the contraction principle (see Lemma 5.7 in [15]) yields $\mathcal{R}_N(\mathcal{F}) \leq \overline{\varphi}_1 \cdot \mathcal{R}_N(\mathcal{F}_1)$. Then, Lemma 1 gives $\mathcal{R}_N(\mathcal{F}_1) \leq \tau_1 \Lambda_1 / \sqrt{N}$ and the result follows.

Assume now that the statement holds for C components, i.e., for $\mathcal{F} = \mathcal{F}^C$:

$$\mathcal{R}_N(\mathcal{F}^C) \leq \sum_{k=1}^C \frac{(\prod_{j=k}^C \overline{\rho}_j) \overline{\varphi}_k \tau_k \Lambda_k}{\sqrt{N}}. \tag{19}$$

Then, we can show that it also holds for $\mathcal{F}^{C+1}$. To see this, first apply the contraction principle again to obtain

$$\begin{aligned}\mathcal{R}_N(\mathcal{F}^{C+1}) &= \mathcal{R}_N(\rho_{C+1} \circ \mathcal{G}_{C+1}(\mathcal{F}^C,\ \varphi_{C+1} \circ \mathcal{F}_{C+1})) \\ &\leq \overline{\rho}_{C+1} \mathcal{R}_N(\mathcal{G}_{C+1}(\mathcal{F}^C,\ \varphi_{C+1} \circ \mathcal{F}_{C+1})).\end{aligned}$$

Then, it remains to show that

$$\begin{aligned}\mathcal{R}_N(\mathcal{G}_{C+1}(\mathcal{F}^C, \varphi_{C+1} \circ \mathcal{F}_{C+1})) \leq \mathcal{R}_N(\mathcal{F}^C) \\ + \mathcal{R}_N(\varphi_{C+1} \circ \mathcal{F}_{C+1})\end{aligned} \tag{20}$$

for all choices of the binary operator g_{C+1} to conclude using (19) and $\mathcal{R}_N(\varphi_{C+1} \circ \mathcal{F}_{C+1}) \leq \overline{\varphi}_{C+1} \cdot \mathcal{R}_N(\mathcal{F}_{C+1})$ with, by Lemma 1, $\mathcal{R}_N(\mathcal{F}_{C+1}) \leq \tau_{C+1} \Lambda_{C+1} / \sqrt{N}$.

If g_{C+1} is the addition, then a basic property of Rademacher complexities (inherited from the subadditivity of the supremum and the linearity of the expectation), namely that $\mathcal{R}_N(\mathcal{U} + \mathcal{V}) \leq \mathcal{R}_N(\mathcal{U}) + \mathcal{R}_N(\mathcal{V})$, suffices to prove (20).

Since the random variables σ_i and $-\sigma_i$ share the same distribution, we have $\mathcal{R}_N(-\mathcal{V}) = \mathcal{R}_N(\mathcal{V})$ and (20) is also proved for the case $\mathcal{G}_{C+1}(\mathcal{U},\mathcal{V}) = \mathcal{U} - \mathcal{V}$.

For the min and max operators, we can follow Lemma 9.1 in [15] and rewrite $g_{C+1}(a,b) = \frac{1}{2}(a+b+s|a-b|)$ with $s=1$ for the $\max$ and $s=-1$ for the $\min$. Then,

$$\mathcal{R}_N(\mathcal{G}_{C+1}(\mathcal{U},\mathcal{V})) \leq \frac{1}{2}\left[\mathcal{R}_N(\mathcal{U}+\mathcal{V}) + \mathcal{R}_N(w\circ(\mathcal{U}-\mathcal{V}))\right]$$

for the function $w : t \mapsto s|t|$ of Lipschitz constant equal to 1. Thus, using the contraction principle again and the results of the two previous cases, we obtain

$$\begin{aligned}\mathcal{R}_N(\mathcal{G}_{C+1}(\mathcal{U},\mathcal{V})) &\leq \frac{1}{2}\left[\mathcal{R}_N(\mathcal{U})+\mathcal{R}_N(\mathcal{V}) + \mathcal{R}_N(\mathcal{U})+\mathcal{R}_N(\mathcal{V})\right] \\ &= \mathcal{R}_N(\mathcal{U}) + \mathcal{R}_N(\mathcal{V})\end{aligned}$$

and thus (20).

Therefore, for any choice of g_{C+1},

$$\begin{aligned}\mathcal{R}_N(\mathcal{F}^{C+1}) &\leq \overline{\rho}_{C+1}\left(\mathcal{R}_N(\mathcal{F}^C) + \mathcal{R}_N(\varphi_{C+1}\circ\mathcal{F}_{C+1})\right) \\ &\leq \overline{\rho}_{C+1}\left(\sum_{k=1}^{C}\frac{(\prod_{j=k}^{C}\overline{\rho}_j)\overline{\varphi}_k\tau_k\Lambda_k}{\sqrt{N}} + \overline{\varphi}_{C+1}\frac{\tau_{C+1}\Lambda_{C+1}}{\sqrt{N}}\right) \\ &= \sum_{k=1}^{C+1}\frac{(\prod_{j=k}^{C+1}\overline{\rho}_j)\overline{\varphi}_k\tau_k\Lambda_k}{\sqrt{N}}\end{aligned}$$

and Theorem 2 is proved by induction for all $C \geq 1$.