Margin theory for the scenario-based approach to robust optimization in high dimension

Fabien Lauer

Margin theory for the scenario-based 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 level of reliability. In this paper, we introduce the notion of margin to improve upon standard results in this field. In particular, using tools from statistical learning theory, we show that the sample complexity of a class of random programs does not explicitly depend on the number of variables. In addition, within the considered class, that includes polynomial constraints among others, this result holds for both convex and nonconvex instances with the same level of guarantees. We also derive a posteriori bounds on the probability of violation and sketch a regularization approach that could be used to improve the reliability of computed solutions on the basis of these bounds.**

***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.

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 random problem is not directly related to its dimension. This stands in contrast to most of the literature. For instance, for convex problems, the sample complexity derived in [2], [4] grows linearly with the dimension. Extensions of these works to nonconvex problems suffer from the same issue. For instance, the approach of [5] relies on decomposing the nonconvex domain as a union of convex ones, and leads to bounds also linear in d for integer programming. Other works [6]–[8] provide bounds on the probability of violation that do not involve d, and instead characterize the complexity in terms of the a posteriori length of the support subsample. However, these bounds only apply after the solution has been computed and cannot be used to estimate the sample complexity. In contrast, we provide a priori bounds on the probability of violation and sample complexity estimates.

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, since there are convex problems for which the probability of violation precisely equals the bound. However, by restricting the class of problems to those with constraints 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 leads to tighter bounds than [2], [4] in the high-dimensional regime. These hold, for instance, for all conjunctions of polynomial constraints for which the uncertain parameters are not involved in the exponents.

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 performance guarantees derived in this paper hold uniformly over the entire domain defined by the constraints without uncertainties. This means that they apply without assumptions on our ability to solve the random scenario program. At the opposite, all approaches [5], [7] that extend the results of [4] to the nonconvex case typically apply only to the global optimizer that satisfies all scenario constraints.

4) A posteriori bounds and regularization: When scenarios cannot be easily generated, a posteriori bounds can be used to assess the quality of a computed solution, as proposed in [6], [8] for convex problems and [7] in the nonconvex case. However, these work rely on the estimation of the support subsample via a rather costly procedure that involves solving multiple optimization problems, which might be computationally demanding in the nonconvex case. Here, we derive a posteriori bounds that can be computed in a straightforward manner for the same cost as a priori bounds. Inspired by these bounds, we describe a regularization procedure that could be used to search for solutions with increased generalization performance.

B. Notations

Sequences are written in bold letters with a subscript indicating their length, e.g., $\boldsymbol{t}_N = (t_i)_{1\leq i\leq N}$. For two vectors $u, v \in \mathbb{R}^n$, the dot product is written as $u^\top v$ and $\|u\|$ denotes the Euclidean norm. However, all results below hold similarly with $\mathbb{R}^n$ replaced by a Hilbert space $\mathcal{H}$ equipped with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ that induces the norm $\|u\|_{\mathcal{H}} = \sqrt{\langle u,u\rangle_{\mathcal{H}}}$. The cardinality of a set $\mathcal{X}$ is denoted by $|\mathcal{X}|$ and, 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 giving the precise formulation of 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. A posteriori bounds are detailed in Sect. III-D and bounds with fast rates of convergence are discussed in Sect. III-E. Finally, optimization problems are considered in Sect. IV and conclusions are given in Sect. V.

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

$$\begin{aligned}&\min_{x\in\mathcal{X}} J(x)\\ &\text{s.t.}\ \ f(x,\theta)\leq 0,\quad \forall\theta\in\Theta.\end{aligned}$$

Here, the domain $\mathcal{X}$ entails all the constraints that are not impacted by uncertainties. Note that no assumption is made regarding its convexity, and $\mathcal{X}$ can typically be a subset of $\mathbb{R}^d$ for continuous optimization, $\mathbb{Z}^d$ for integer programming or of $\mathbb{R}^{d_1}\times\mathbb{Z}^{d_2}$ for mixed-integer programs.

While the proposed approach described in this paper 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 $\eta_k : \Theta \to \mathbb{R}$. Note that no assumption will be made regarding the convexity of these functions and of the resulting optimization problems. 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). More generally, we will 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 $f = f^C$, $f^1 = \varphi_1 \circ f_1$, and, for $k\geq 2$, $f^k = \rho_k \circ g_k(f^{k-1}, \varphi_k \circ f_k)$.

This formulation 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 $\max\{\min\{f_1(x,\theta), f_2(x,\theta)\}, f_3(x,\theta)\}\leq 0$. As another example, the function $f(x,\theta) =$

$\sqrt{1.1+\sin(\theta x_1 x_2)} - \cos(\theta x_2^3)$ can also be written as (4). Polynomial constraints could be implemented as well as sums of C monomials $f_k(x,\theta)$. However, it will be more efficient to consider instead $C = 1$ and let $f_1(x,\theta)$ be a polynomial with $\phi_1(x)$ gathering all its monomials and $\psi_1(\theta)$ its coefficients.[1]

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 $n_1 = \cdots = n_C$, 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

$$\begin{aligned} \hat{x} = &\operatorname*{argmin}_{x\in\mathcal{X}} J(x) \\ \text{s.t. } &f(x,\theta_i) \leq 0, \quad i = 1,\dots,N. \end{aligned}$$

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 for instance that for *convex* problems and $\hat{x}$ the solution to Problem 2,

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

where

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

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.

[1]While the dimension n_1 of $\phi_1(x)$ could quickly grow with the dimension d of x or the degree of the polynomial, we should only keep the monomials with uncertainty in their coefficients as components of $\phi_1(x)$ and then append another single component to compute the sum of all the remaining terms with the corresponding entry in $\psi_1(\theta)$ set to 1.

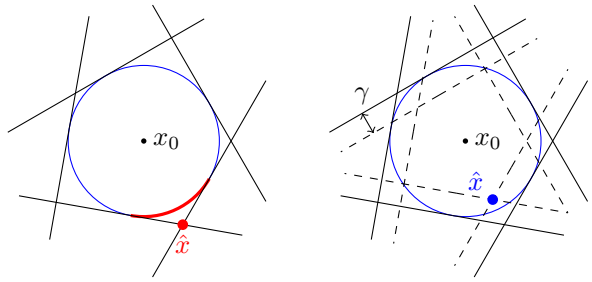

Fig. 1. Illustration of the benefit of the margin on a toy example: the constraint $\|x - x_0\| \leq 1$ (blue circle) is implemented by an infinite number of linear constraints, $f(x,\theta) = \theta^\top(x - x_0) - 1 \leq 0$, for θ uniformly distributed in the unit circle. *Left:* The standard scenario 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 $\hat{x}$, satisfy the 5 scenario constraints with a margin γ (the 5 dashed lines) and also satisfy the worst-case constraint (blue circle) with $V(\hat{x}) = 0$.

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

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

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.

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.

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

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

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.

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 $\ell_\gamma : \mathcal{X}\times\Theta \to [0,1]$

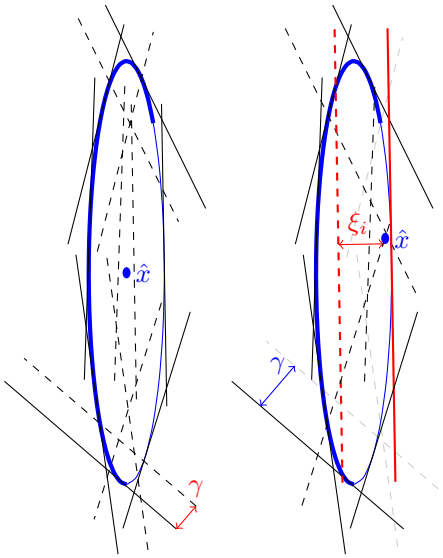

Fig. 2. Illustration of the soft-margin approach on a toy example where random linear constraints are substituted for a nonlinear constraint (blue ellipsoid). *Left:* The hard-margin scenario program yields a solution $\hat{x}$ that is feasible for all $\theta \in \Theta$ with $V(\hat{x}) = 0$. However, the margin γ is rather small and cannot be increased without making Problem 3 infeasible. Thus, the upper bound on $V(\hat{x})$ given by Theorem 1 is not really tight and, though the solution is perfect, we have a 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 $\hat{V}_\gamma(\hat{x})$.

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

$$\hat{V}_\gamma(x) = \frac{1}{N}\sum_{i=1}^{N} \ell_\gamma(x,\theta_i).$$

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 $V(x)$, and this can be done at a finer scale 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 6 in Appendix A 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 (Lemma 1 in App. A) 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

$$\mathcal{R}_N(\mathcal{L}_\gamma) \leq \frac{1}{\gamma}\mathcal{R}_N(\mathcal{F})$$

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., [13]): 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 $\hat{x}$ lies in the domain $\mathcal{X}$. This is all the more relevant when considering nonconvex instances for which true solutions are difficult to obtain, and stands in contrast to the standard results discussed in the literature [4], [5], [7] that only apply to the optimizer of the scenario problem.

The influence of the margin γ on the performance guarantees are 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 $\hat{V}_\gamma(x)$ increases by $\xi_i/\gamma N$ as one of the margin constraints is violated.[2] 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 shows how to upper bound this complexity for the specific case of constraints 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.

Proof: See Appendix B. □

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 the size of the $\phi_k(x)$'s over $\mathcal{X}$ (as measured by Λ_k).[3] This stands in contrast to most of the literature on scenario optimization, such as [3]–[5], 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], the VC-dimension d_{VC} in (14) grows linearly with d even for the most simple constraint function (4) with $C = 1$, φ_1 the identity and f_1 a linear function of x. Finally, when considering integer programming problems, the method proposed in [5] also leads to a bound linear in the number of integer variables.

However, there is nothing special about vectors $\phi_k(x)$ with small norms, i.e., closely gathered around the origin. Indeed, 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 of the form (4). Then, for any choice 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 2, 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

$$\sup_{x\in\mathcal{X}}\|\tilde{\phi}_k(x)\| = \sup_{x\in\mathcal{X}}\|\phi_k(x) - \phi_{0,k}\| = \tilde{\Lambda}_k.$$

□

Note that Corollary 1 provides an improvement over Theorem 2, since it allows the choice $\phi_{0,k} = 0$.

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, at first glance, might appear weaker than most results in the literature on scenario optimization, such as (7)–(8), which typically lead to bounds with a faster convergence rate 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 is all the more relevant that scenario optimization is mostly concerned with the non-asymptotic case which can offer practical solutions to difficult robust optimization problems with a finite and reasonable N, thus limiting the benefit of bounds in $O(d/N)$ to rather

[2]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, $\xi_i < \gamma$.

[3]Also note that, due to the definition of Λ_k, $\mathcal{X}$ could be replaced by any (possibly discrete) subset $\mathcal{X}' \subset \mathcal{X}$ without any negative impact on the statistical guarantees.

small dimensions in practice. 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).[4]

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.

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

$$N(\epsilon,\delta) = \frac{\left(\frac{2}{\gamma}\sum_{k=1}^{C}\tau_k\Lambda_k + \sqrt{\log\frac{1}{\delta^{1/2}}}\right)^2}{\epsilon^2}, \qquad (15)$$

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 $\hat{x}$ satisfy $\hat{V}_\gamma(\hat{x}) = 0$ and the bound $V(\hat{x}) \leq \epsilon$ 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

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

which grows linearly with the dimension d. Therefore, it suffices that

$$d > \frac{\left(\frac{2}{\gamma}\sum_{k=1}^{C}\tau_k\Lambda_k + \sqrt{\log\frac{1}{\delta^{1/2}}}\right)^2}{2\epsilon}$$

to observe $N_{\text{cvx}}(\epsilon,\delta) > N(\epsilon,\delta)$. Consider for instance convex problems with $C=1$, $\tau_1\Lambda_1/\gamma = 2$, $\epsilon = 0.03$ and $\delta = 0.001$. Then, $N_{\text{cvx}}(\epsilon,\delta) > N(\epsilon,\delta)$ and the proposed approach improves upon [4] for all $d \geq 573$.

C. Optimizing 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 the following.

Problem 5 (Max-margin scenario program): Given a probability distribution of $\theta \in \Theta$, draw a sample of N independent parameter values $(\theta_i)_{1\leq i\leq N} \subset \Theta$ and solve

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

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

However, Theorem 1 only holds for a value of γ fixed a priori and cannot be applied with the margin resulting from Problem 5 and that depends on the random data $(\theta_i)_{1\leq i\leq N}$. Nontetheless, it can be extended to hold uniformly over γ by following the proof of Theorem 5.9 in [13], with the addition of a small term in $O(\sqrt{\log\log_2(1/\gamma)})$.

Theorem 3: Given an upper bound $\overline{\gamma} > 0$ on the margin 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

$$V(x) \leq V_\gamma(x) \leq \inf_{\gamma\in(0,\overline{\gamma}]} \hat{V}_\gamma(x) + \frac{4}{\gamma}\mathcal{R}_N(\mathcal{F}) + \sqrt{\frac{\log\frac{1}{\delta}}{2N}} + \sqrt{\frac{\log\log_2\frac{2\overline{\gamma}}{\gamma}}{N}},$$

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

Therefore, the value of γ could be chosen a posteriori, in order to produce the best bound on the probability of violation. This could be done also with the standard scenario program of Problem 2 that returns a feasible point $\hat{x}$, possibly within the margin for some of the constraints. In this case, increasing the margin decreases the confidence interval of Theorem 3, while it increases the empirical margin risk $\hat{V}_\gamma(x)$, and the best trade-off between these two terms can be sought for.

D. A posteriori bounds and regularization

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 on the basis of 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 that measure quantities computed for the specific solution $\hat{x}$ returned by the algorithm rather than global upper bounds. All these modifications are summarized in the following theorem.

Theorem 4: Let f be of the form (4). Then, for any choice of centers $(\phi_{0,k})_{1\leq k\leq C} \in \prod_{k=1}^{C}\mathbb{R}^{n_k}$ 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

$$V(x) \leq \hat{V}_\gamma(x) + \frac{2\overline{\Lambda}(x)\sum_{k=1}^{C}\left(\prod_{j=k}^{C}\overline{\rho}_j\right)\overline{\varphi}_k\sqrt{\sum_{i=1}^{N}\|\psi_k(\theta_i)\|^2}}{\gamma N} + 3\sqrt{\frac{\log\frac{6}{\delta}}{2N}} + 3\sqrt{\frac{\log\log_2 2\overline{\Lambda}(x)}{N}}$$

with $\overline{\Lambda}(x) = \max\{1, \|\phi_1(x) - \phi_{0,1}\|, \ldots, \|\phi_C(x) - \phi_{0,C}\|\}$.

Proof: See Appendix C. □

With Theorem 4 at hand, it is now possible to envision a regularized scenario approach, in which we compute $\hat{x}$ as the solution that minimizes $\overline{\Lambda}(x)$. According to Theorem 4, this $\hat{x}$ should enjoy better generalization performance than others with similar empirical error $\hat{V}_\gamma(x)$. In particular, if we compute $\hat{x}$ using the hard margin constraints of Problem 3, then regularization offers a sound tie-breaking rule to choose among all feasible x with $\hat{V}_\gamma(x) = 0$.

In practice, the centers $(\phi_{0,k})_{1\leq k\leq C}$ appearing in $\overline{\Lambda}(x)$ must be fixed in advance and can typically be chosen as the centers of the sets $\{\phi_k(x) : x \in \mathcal{X}\}$. Alternatively, one could compute the centers as $\phi_{0,k} = \phi_k(\hat{x})$ for a preliminary guess $\hat{x}$ of the solution obtained on an independent sample of scenarios.

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 [14], [15] to yield bounds in $O(1/N)$ that are still independent of the dimension d of x. However, we retain from giving the details for two reasons. First, such extensions typically apply only to the minimizer $\hat{x}$ of $\hat{V}_\gamma(x)$, which can be difficult to compute in nonconvex cases. Second, this also introduces rather large constants and the interest of the resulting bounds remains questionable in the non-asymptotic case (say, for $N < 10^7$). In statistical learning, another path that leads to fast rate bounds is based on the use of covering numbers instead of the Rademacher complexity and relies on relative deviation bounds, as in [16]. Following these ideas, we obtain the risk bound below for the margin-based scenario approach.[5]

Theorem 5: Let the constraints be of the form (4). Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ on the random draw of $(\theta_i)_{1\leq i\leq N}$, the probability of violation is bounded, for any feasible x in Problem 3, by

$$V(x) \leq 4\frac{144\sum_{k=1}^{C} M_k^2 \log(60 M_k N) + \log\frac{4}{\delta}}{N},$$

where

$$M_k = \frac{2^{p_k}\overline{\varphi}_k\tau_k\Lambda_k\prod_{j=k}^{C}\overline{\rho}_j}{\gamma},$$

[5]For the sake of clarity, we only expose the result for the hard-margin version (or zero-error case) here, but a more general bound including errors is proved in Appendix D.

p_k is the number of binary operations g_j with $j = k, \ldots, C$ that are sums or differences, and all other constants are as in Theorem 2.

Proof: See Appendix D. □

Again, this result does not explicitly depend on d, and shows an improved convergence rate of $O(\log(N)/N)$. This can be transformed into a sample complexity estimate that grows in $O(\log(\epsilon)/\epsilon)$. Yet, the constants appearing in the above are larger than in the previous results based on the Rademacher complexity, which makes the improvement in terms of the convergence rate visible mostly for $N > 10^5$.

IV. Optimization problems and margin complexity

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

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

However, introducing the margin can affect the objective value $J(\hat{x})$ of the computed solution $\hat{x}$. 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 by tuning $\lambda > 0$ and solving

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

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 $J(\hat{x})$ will result from the smallest choice of margin γ in (17). 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 (17), 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 (17) with $\gamma = \gamma(N,\epsilon,\delta)$ given above. Then, we have the guarantee that the solution $\hat{x}$ satisfies $P^N\{V(\hat{x}) \leq \epsilon\} \geq 1-\delta$.

V. 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 dependence on the problem dimension could be removed. These results hold only for a specific class of constraint functions. But this class includes polynomials as particular cases and subsumes typical classes previously

considered in the literature. In addition, the results derived in this paper 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 investigated further. More generally, this paper also encourages the use of techniques from learning theory that have been overlooked so far in the context of scenario optimization.

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

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

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

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

[17] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin, 1991.

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

[19] F. Lauer, "Error bounds for piecewise smooth and switching regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 4, pp. 1183–1195, 2020.

APPENDIX A
TOOLS FOR RADEMACHER COMPLEXITIES

The following general theorem is a classical tool from statistical learning theory, see, e.g., Theorem 3.3 in [13].

Theorem 6: Let $\mathcal{L}$ be a class of functions from Θ into $[0,1]$ and $(\theta_i)_{1\leq i\leq N}$ be a sequence of independent copies of the random variable $\theta \in \Theta$. Then, for a fixed $\delta \in (0,1)$, with probability at least $1-\delta$, each of the following holds uniformly over all $\ell \in \mathcal{L}$,

$$\mathbb{E}_\theta \ell(\theta) \leq \frac{1}{N}\sum_{i=1}^N \ell(\theta_i) + 2\mathcal{R}_N(\mathcal{L}) + \sqrt{\frac{\log\frac{1}{\delta}}{2N}},$$

$$\mathbb{E}_\theta \ell(\theta) \leq \frac{1}{N}\sum_{i=1}^N \ell(\theta_i) + 2\hat{\mathcal{R}}_N(\mathcal{L}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2N}}.$$

We recall the contraction principle for Rademacher complexities, found in Theorem 4.12 of [17] or Lemma 5.7 of [13].

Lemma 1: If the function $\varphi : \mathbb{R} \to \mathbb{R}$ is $\overline{\varphi}$-Lipschitz, i.e., if $\forall (u,v) \in \mathbb{R}^2,\ |\varphi(u) - \varphi(v)| \leq \overline{\varphi}|u-v|$, then, for the class $\varphi \circ \mathcal{F} = \{\varphi \circ f : f \in \mathcal{F}\}$,

$$\hat{\mathcal{R}}_N(\varphi \circ \mathcal{F}) \leq \overline{\varphi} \cdot \hat{\mathcal{R}}_N(\mathcal{F}).$$

APPENDIX B
PROOF OF THEOREM 2

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 2: Given functions $\psi : \Theta \times \mathbb{R}^n$, $\phi : \Theta \to \mathbb{R}^n$ 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

$$\begin{aligned}\mathbb{E}\sup_{x\in\mathcal{X}} \frac{1}{N}\sum_{i=1}^N \sigma_i\eta(\theta_i) &= \mathbb{E}\frac{1}{N}\sum_{i=1}^N \sigma_i\eta(\theta_i) \\ &= \frac{1}{N}\sum_{i=1}^N \eta(\theta_i)\mathbb{E}\sigma_i = 0,\end{aligned}$$

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

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

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:

$$\begin{aligned}\mathcal{F} &= \mathcal{F}^C\\ \mathcal{F}^1 &= \varphi_1\circ\mathcal{F}_1\\ \mathcal{F}^{k+1} &= \rho_{k+1}\circ\mathcal{G}_{k+1}(\mathcal{F}^k,\ \varphi_{k+1}\circ\mathcal{F}_{k+1}).\end{aligned} \tag{18}$$

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 of Lemma 1 in Appendix A yields

$$\mathcal{R}_N(\mathcal{F}) \leq \overline{\varphi}_1\cdot\mathcal{R}_N(\mathcal{F}_1).$$

Then, Lemma 2 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 2, $\mathcal{R}_N(\mathcal{F}_{C+1}) \leq \tau_{C+1}\Lambda_{C+1}/\sqrt{N}$.

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 [13] 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

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

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 $C \geq 1$.

APPENDIX C
PROOF OF THEOREM 4

Define

$$R = \sum_{k=1}^{C} \left(\prod_{j=k}^{C} \overline{\rho}_j \right) \overline{\varphi}_k \sqrt{\sum_{i=1}^{N} \|\psi_k(\theta_i)\|^2}$$

and, for any $\alpha > 9/2$, let $p = 2\alpha/9$, $\zeta(p)$ denote the Riemann zeta function and

$$\epsilon = 3\sqrt{\frac{\log \frac{2+2\zeta(p)}{\delta}}{2N}}, \tag{21}$$

$$\epsilon_x = \epsilon + \sqrt{\frac{\alpha \log \log_2 2\overline{\Lambda}(x)}{N}}.$$

We will show that, for any $\delta \in (0,1)$,

$$P^N \left\{ \forall x \in \mathcal{X},\ V_\gamma(x) \leq \hat{V}_\gamma(x) + \frac{4\overline{\Lambda}(x)R}{\gamma N} + \epsilon_x \right\} \geq 1 - \delta, \tag{22}$$

from which Theorem 4 is derived by choosing $\alpha = 9$ and noting that $\zeta(2) = \pi^2/6 < 2$.

For any subset $\mathcal{X}_j \subseteq \mathcal{X}$, let $\mathcal{F}(\mathcal{X}_j) = \{f_x \in \mathbb{R}^\Theta : f_x(\theta) = f(x,\theta),\ x \in \mathcal{X}_j\}$. Following the proof of Theorem 1 while using the second inequality of Theorem 6 instead of the first yields

$$P^N \left\{ \exists x \in \mathcal{X}_j, V_\gamma(x) > \hat{V}_\gamma(x) + \frac{2}{\gamma}\hat{\mathcal{R}}_N(\mathcal{F}(\mathcal{X}_j)) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2N}} \right\} \leq \delta. \tag{23}$$

Then, Theorem 2 and Corollary 1 can be reshaped to yield

$$\hat{\mathcal{R}}_N(\mathcal{F}(\mathcal{X}_j)) \leq \sum_{k=1}^{C} \left(\prod_{k'=k}^{C} \overline{\rho}_{k'} \right) \overline{\varphi}_k \sup_{x \in \mathcal{X}_j} \|\phi_k(x) - \phi_{0,k}\| \times \frac{\sqrt{\sum_{i=1}^{N} \|\psi_k(\theta_i)\|^2}}{N}.. \tag{24}$$

To see this, note that their proofs hold verbatim with the empirical version of the Rademacher complexities and all $\tau_k/\sqrt{N}$ replaced by $\sqrt{\sum_{i=1}^{N} \|\psi_k(\theta_i)\|^2/N}$, as given by Lemma 2.

Now, let $l_j = 2^j$ and, for any integer $j \geq 1$, $\epsilon_j = \epsilon + \sqrt{\alpha \log \log_2(l_j)/N}$ and

$$\mathcal{X}_j = \{x \in \mathcal{X} : \Lambda(x) = \max_{k \in \{1,\ldots,C\}} \|\phi_k(x) - \phi_{0,k}(x)\| \leq l_j\} \tag{25}$$

with $\overline{\Lambda}(x) = \max\{1, \Lambda(x)\}$. For any $x \in \mathcal{X}$, either $x \in \mathcal{X}_0$ with $\overline{\Lambda}(x) = 1 > 1/2$ and $\epsilon_x = \epsilon$, or there is a $j \geq 1$ such that $x \in \mathcal{X}_j$ with $\overline{\Lambda}(x) = \Lambda(x)$, $l_j \geq \Lambda(x) > l_{j-1} = l_j/2$ and $\epsilon_x > \epsilon_j$. Thus,

$$\begin{aligned}
&P^N \left\{ \exists x \in \mathcal{X},\ V_\gamma(x) > \hat{V}_\gamma(x) + \frac{4\overline{\Lambda}(x)R}{\gamma N} + \epsilon_x \right\} \\
&= P^N \left\{ \exists j \geq 0, x \in \mathcal{X}_j, V_\gamma(x) > \hat{V}_\gamma(x) + \frac{4\overline{\Lambda}(x)R}{\gamma N} + \epsilon_x \right\} \\
&\leq P^N \left\{ x \in \mathcal{X}_0,\ V_\gamma(x) > \hat{V}_\gamma(x) + \frac{2R}{\gamma N} + \epsilon \right\} \\
&\quad + P^N \left\{ \exists j \geq 1, x \in \mathcal{X}_j, V_\gamma(x) > \hat{V}_\gamma(x) + \frac{2l_jR}{\gamma N} + \epsilon_j \right\}.
\end{aligned}$$

For $\mathcal{X}_j$ as in (25), the bound (24) yields

$$\hat{\mathcal{R}}_N(\mathcal{F}(\mathcal{X}_j)) \leq \frac{l_j R}{N}$$

and thus (23) gives, for $j = 0$,

$$P^N \left\{ x \in \mathcal{X}_0,\ V_\gamma(x) > \hat{V}_\gamma(x) + \frac{2R}{\gamma N} + \epsilon \right\} \leq 2e^{-2N\epsilon^2/9}$$

and, for any $j \geq 1$,

$$P^N \left\{ x \in \mathcal{X}_j,\ V_\gamma(x) > \hat{V}_\gamma(x) + \frac{2l_jR}{\gamma N} + \epsilon_j \right\} \leq 2e^{-2N\epsilon_j^2/9}.$$

Since $l_j = 2^j$, we have

$$\epsilon_j^2 = \left(\epsilon + \sqrt{\frac{\alpha \log j}{N}} \right)^2 \geq \epsilon^2 + \frac{\alpha \log j}{N}.$$

Thus, with $p = 2\alpha/9$,

$$\begin{aligned}
e^{-2N\epsilon_j^2/9} &\leq e^{-2N(\epsilon^2 + \alpha \log(j)/N)/9} \\
&= e^{-2N\epsilon^2/9} e^{-p \log(j)} = \frac{e^{-2N\epsilon^2/9}}{j^p}.
\end{aligned}$$

Applyng the union bound, this leads to

$$\begin{aligned}
&P^N \left\{ \exists j \geq 1, x \in \mathcal{X}_j,\ V_\gamma(x) > \hat{V}_\gamma(x) + \frac{4\overline{\Lambda}(x)R}{\gamma N} + \epsilon_x \right\} \\
&\leq 2 \sum_{j \geq 1} e^{-2N\epsilon_j^2/9} \\
&\leq 2e^{-2N\epsilon^2/9} \sum_{j \geq 1} \frac{1}{j^p} = 2\ \zeta(p) e^{-2N\epsilon^2/9}.
\end{aligned}$$

Therefore,

$$P^N \left\{ \exists x \in \mathcal{X},\ V_\gamma(x) > \hat{V}_\gamma(x) + \frac{4\overline{\Lambda}(x)R}{\gamma N} + \epsilon_x \right\} \leq (2 + 2\zeta(p)) e^{-2N\epsilon^2/9},$$

which, by recalling the value of ϵ in (21), is precisely (22).

APPENDIX D
PROOF OF THEOREM 5

Throughout this section, we consider a slightly different margin loss function:

$$\ell_\gamma(x,\theta) = \mathbf{1}_{f(x,\theta) > -\gamma} \geq \mathbf{1}_{f(x,\theta) > 0}. \tag{26}$$

We will first prove the following more general result: with probability at least $1-\delta$, for all $x \in \mathcal{X}$,

$$V(x) \leq V_\gamma(x) \leq \hat{V}_\gamma(x) + 2\sqrt{\hat{V}_\gamma(x)\frac{M+\log\frac{4}{\delta}}{N}} + 4\frac{M+\log\frac{4}{\delta}}{N} \tag{27}$$

with $V_\gamma(x)$ and $\hat{V}_\gamma(x)$ computed using (26) and

$$M = 144\sum_{k=1}^{C} M_k^2 \log(60 M_k N)$$

with

$$M_k = \frac{2^{p_k}\overline{\varphi}_k \tau_k \Lambda_k \prod_{j=k}^{C}\overline{\rho}_j}{\gamma}.$$

Then, Theorem 5 will follow by assuming that $f(x,\theta_i) \leq -\gamma$, $i = 1,\dots,N$, which ensures that $\hat{V}_\gamma(x) = 0$ in (27).

This result relies on the following capacity measure.

Definition 2 (ϵ-nets and covering numbers): Given a function class $\mathcal{F} \subset \mathbb{R}^\Theta$ and a sequence $\boldsymbol{\theta}_N \in \Theta^N$, consider the empirical pseudo-metric defined by

$$\forall (f,f') \in \left(\mathbb{R}^\Theta\right)^2, \quad d_{\boldsymbol{\theta}_N}(f,f') = \max_{1\leq i\leq N} |f(\theta_i) - f'(\theta_i)|.$$

Then, an *ϵ-net* of $\mathcal{F}$ is any set $\mathcal{F}' \subset \mathbb{R}^\Theta$ such that $\forall f \in \mathcal{F}$, $d_{\boldsymbol{\theta}_N}(f,\mathcal{F}') = \min_{f'\in\mathcal{F}'} d_{\boldsymbol{\theta}_N}(f,f') < \epsilon$ and the *covering number* $\mathcal{N}(\epsilon,\mathcal{F},\boldsymbol{\theta}_N)$ at scale ϵ of $\mathcal{F}$ is the cardinality of its smallest ϵ-net. *Uniform covering numbers* are defined by

$$\mathcal{N}(\epsilon,\mathcal{F},N) = \sup_{\boldsymbol{\theta}_N\in\Theta^N} \mathcal{N}(\epsilon,\mathcal{F},\boldsymbol{\theta}_N).$$

We will also require the following contraction lemma.

Lemma 3: Let $\varphi : R \to \mathbb{R}$ be a $\overline{\varphi}$-Lipschitz continuous function over its domain $R \subset \mathbb{R}$, i.e., $\forall u,v \in R$, $|\varphi(u) - \varphi(v)| \leq \overline{\varphi}|u-v|$. Then, for all $\epsilon > 0$,

$$\mathcal{N}(\epsilon,\varphi\circ\mathcal{F},N) \leq \mathcal{N}\left(\frac{\epsilon}{\overline{\varphi}},\mathcal{F},N\right).$$

Proof: For any $\boldsymbol{\theta}_N = (\theta_i)_{1\leq i\leq N}$, let $\hat{\mathcal{F}}$ be a minimal $\epsilon/\overline{\varphi}$-cover of $\mathcal{F}$ with cardinality $\mathcal{N}(\epsilon/\overline{\varphi},\mathcal{F},\boldsymbol{\theta}_N)$. Then, for any $f \in \mathcal{F}$, there is a $\hat{f} \in \hat{\mathcal{F}}$ such that

$$\begin{aligned} d_{\boldsymbol{\theta}_N}(\varphi\circ f, \varphi\circ\hat{f}) &= \max_{1\leq i\leq N} |\varphi(f(\theta_i)) - \varphi(\hat{f}(\theta_i))| \\ &\leq \max_{1\leq i\leq N} \overline{\varphi}|f(\theta_i) - \hat{f}(\theta_i)| \\ &\leq \overline{\varphi} d_{\boldsymbol{\theta}_N}(f,\hat{f}) < \epsilon \end{aligned}$$

and $\varphi\circ\hat{\mathcal{F}}$ is an ϵ-net of $\varphi\circ\mathcal{F}$ of cardinality $\mathcal{N}(\epsilon/\overline{\varphi},\mathcal{F},\boldsymbol{\theta}_N)$. Taking the supremum over all $\boldsymbol{\theta}_N$ concludes the proof. □

The following lemmas will provide bounds on the covering numbers of $\varphi_k \circ \mathcal{F}_k$ and of the combination of two classes.

Lemma 4: Given functions $\psi : \Theta \times \mathbb{R}^n$, $\phi : \Theta \to \mathbb{R}^n$, $\eta : \Theta \to \mathbb{R}$ and a $\overline{\varphi}$-Lipschitz continuous function $\varphi : \mathbb{R} \to \mathbb{R}$, the uniform covering numbers of the class $\mathcal{F} = \{f_x \in \mathbb{R}^\Theta : f_x(\theta) = \varphi\left(\psi(\theta)^\top\phi(x) + \eta(\theta)\right),\ x \in \mathcal{X}\}$ are bounded, for all $\epsilon \in (0, \tau\Lambda]$, by

$$\log\mathcal{N}(\epsilon,\mathcal{F},N) \leq \frac{36\overline{\varphi}^2\tau^2\Lambda^2}{\epsilon^2}\log\frac{15\overline{\varphi}\tau\Lambda N}{\epsilon},$$

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

Proof: Let $\tilde{\mathcal{F}} = \{f_x \in \mathbb{R}^\Theta : f_x(\theta) = \psi(\theta)^\top\phi(x) + \eta(\theta),\ x \in \mathcal{X}\}$ such that $\mathcal{F} = \varphi\circ\tilde{\mathcal{F}}$. Then, the result follows from Lemma 3 and the following bound on the covering numbers of $\tilde{\mathcal{F}}$.

For all $\epsilon \in (0,\tau\Lambda]$, Theorem 4 in [18] yields an ϵ-net of $\mathcal{F}' = \{f'_x \in \mathbb{R}^\Theta : f'_x(\theta) = \psi(\theta)^\top\phi(x),\ x \in \mathcal{X}\}$ with cardinality n_1 such that

$$\begin{aligned} \log n_1 &\leq \frac{36\tau^2\Lambda^2}{\epsilon^2}\log\left(2N\left\lceil\frac{4\tau\Lambda}{\epsilon}+2\right\rceil+1\right) \\ &< \frac{36\tau^2\Lambda^2}{\epsilon^2}\log\frac{15\tau\Lambda N}{\epsilon}. \end{aligned}$$

For all $\hat{f}'_x$ in this ϵ-net of $\mathcal{F}'$, let $\hat{f}_x = \hat{f}'_x + \eta$ and note that any function $f_x \in \tilde{\mathcal{F}}$ can be decomposed as $f_x = f'_x + \eta$. Thus, for any $f_x \in \tilde{\mathcal{F}}$, there is an $\hat{f}_x$ such that

$$\begin{aligned} d(f_x,\hat{f}_x) &= \max_{1\leq i\leq N} |f'_x(\theta_i) + \eta(\theta_i) - \hat{f}'_x(\theta_i) - \eta(\theta_i)| \\ &= \max_{1\leq i\leq N} |f'_x(\theta_i) - \hat{f}'_x(\theta_i)| < \epsilon \end{aligned}$$

and the set of functions $\hat{f}_x$ thus built forms an ϵ-net of $\tilde{\mathcal{F}}$. Since the cardinality of this set is n_1, we conclude that $\mathcal{N}(\epsilon,\tilde{\mathcal{F}},N) \leq n_1$.

□

Lemma 5: Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be subsets of $\mathbb{R}^\Theta$ and $\mathcal{G}(\mathcal{F}_1,\mathcal{F}_2)$ be either $\mathcal{F}_1+\mathcal{F}_2$, $\mathcal{F}_1-\mathcal{F}_2$, $\max(\mathcal{F}_1,\mathcal{F}_2)$ or $\min(\mathcal{F}_1,\mathcal{F}_2)$. Then, for any $\epsilon > 0$,

$$\mathcal{N}(\epsilon,\mathcal{G}(\mathcal{F}_1,\mathcal{F}_2),N) \leq \mathcal{N}\left(\frac{\epsilon}{2^a},\mathcal{F}_1,N\right)\mathcal{N}\left(\frac{\epsilon}{2^a},\mathcal{F}_2,N\right),$$

where a is 1 if $\mathcal{G} = \mathcal{F}_1+\mathcal{F}_2$ or $\mathcal{G} = \mathcal{F}_1-\mathcal{F}_2$ and 0 otherwise.

Proof: Let $\hat{\mathcal{F}}_1$ and $\hat{\mathcal{F}}_2$ denote minimal ϵ-nets of $\mathcal{F}_1$ and $\mathcal{F}_2$ of cardinality $\mathcal{N}(\epsilon,\mathcal{F}_1,N)$ and $\mathcal{N}(\epsilon,\mathcal{F}_2,N)$, respectively. Then, for any $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$, there are $\hat{f}_1 \in \hat{\mathcal{F}}_1$ and $\hat{f}_2 \in \hat{\mathcal{F}}_2$ such that

$$\begin{aligned} &d_{\boldsymbol{\theta}_N}(f_1+f_2,\ \hat{f}_1+\hat{f}_2) \\ &\quad= \max_{1\leq i\leq N} |f_1(\theta_i) + f_2(\theta_i) - \hat{f}_1(\theta_i) - \hat{f}_2(\theta_i)| \\ &\quad\leq \max_{1\leq i\leq N} |f_1(\theta_i) - \hat{f}_1(\theta_i)| + |f_2(\theta_i) - \hat{f}_2(\theta_i)| \\ &\quad\leq d_{\boldsymbol{\theta}_N}(f_1,\hat{f}_1) + d_{\boldsymbol{\theta}_N}(f_2,\hat{f}_2) < 2\epsilon. \end{aligned}$$

Thus, $\{\hat{f}_1+\hat{f}_2,\ \hat{f}_1\in\hat{\mathcal{F}}_1, \hat{f}_2\in\hat{\mathcal{F}}_2\}$ is a (2ϵ)-net of $\mathcal{F}_1+\mathcal{F}_2$ of cardinality no larger than $|\hat{\mathcal{F}}_1|\cdot|\hat{\mathcal{F}}_2|$, and the result follows by rescaling ϵ and setting $a = 1$.

The same argument yields $d_{\boldsymbol{\theta}_N}(f_1-f_2,\hat{f}_1-\hat{f}_2) < 2\epsilon$, and the result for $\mathcal{F}_1-\mathcal{F}_2$ with again $a=1$. The proof for $\max(\mathcal{F}_1,\mathcal{F}_2)$ and $\min(\mathcal{F}_1,\mathcal{F}_2)$ can be found in Lemma 13 of [19], where the scale ϵ is not affected by the operation and a can be set to 0.

□

We are now ready to formulate the bound on the covering numbers of the class of functions f as in (4), which will play a role similar to Theorem 2 in the proof of (27).

Lemma 6: Let $\mathcal{F}$ be a function class as in (11) with f of the form (4). Then,

$$\log \mathcal{N}(\epsilon, \mathcal{F}, N) \leq \frac{36}{\epsilon^2} \sum_{k=1}^{C} 4^{p_k} \left(\prod_{j=k}^{C} \overline{\rho}_j \right)^2 \overline{\varphi}_k^2 \tau_k^2 \Lambda_k^2 \times \log \frac{15 \cdot 2^{p_k} (\prod_{j=k}^{C} \overline{\rho}_j) \overline{\varphi}_k \tau_k \Lambda_k N}{\epsilon},$$

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, and $p_k = \sum_{j=k}^{C} a_j$ with $a_j = 1$ if the binary operation g_j is a sum or a difference and 0 otherwise.

Proof: Consider the notations defined in (18). For $C = 1$, $\mathcal{F} = \varphi_1 \circ \mathcal{F}_1$ and Lemma 4 yields the desired result (note that p_1 is always zero in this case).

Assume now that the statement holds for $\mathcal{F}^C$. Then, Lemma 3 yields

$$\log \mathcal{N}(\epsilon, \mathcal{F}^{C+1}, N) \leq \log \mathcal{N}\left(\frac{\epsilon}{\overline{\rho}_{C+1}}, \mathcal{G}_{C+1}(\mathcal{F}^C, \varphi_{C+1} \circ \mathcal{F}_{C+1}), N \right).$$

Let $\epsilon' = \epsilon / \overline{\rho}_{C+1}$. Then, Lemma 5 gives

$$\log \mathcal{N}(\epsilon', \mathcal{G}_{C+1}(\mathcal{F}^C, \varphi_{C+1} \circ \mathcal{F}_{C+1}), N) \leq \log \mathcal{N}\left(\frac{\epsilon'}{2^{a_{C+1}}}, \mathcal{F}^C, N \right) + \log \mathcal{N}\left(\frac{\epsilon'}{2^{a_{C+1}}}, \varphi_{C+1} \circ \mathcal{F}_{C+1}, N \right),$$

with $a_{C+1} = 0$ if g_{C+1} is the maximum or minimum operator and $a_{C+1} = 1$ if g_k is a sum or a difference. Using this with the assumption on the covering numbers of $\mathcal{F}^C$ and Lemma 4 further leads to

$$\begin{aligned}
&\log \mathcal{N}(\epsilon', \mathcal{G}_{C+1}(\mathcal{F}^C, \varphi_{C+1} \circ \mathcal{F}_{C+1}), N) \\
&\quad \leq \frac{36 \cdot 4^{a_{C+1}}}{\epsilon'^2} \sum_{k=1}^{C} 4^{p_k} \left(\prod_{j=k}^{C} \overline{\rho}_j \right)^2 \overline{\varphi}_k^2 \tau_k^2 \Lambda_k^2 \times \log \frac{15 \cdot 2^{a_{C+1}} \cdot 2^{p_k} (\prod_{j=k}^{C} \overline{\rho}_j) \overline{\varphi}_k \tau_k \Lambda_k N}{\epsilon'} \\
&\qquad + \frac{36 \cdot 4^{a_{C+1}} \overline{\rho}_{C+1}^2 \overline{\varphi}_{C+1}^2 \tau_{C+1}^2 \Lambda_{C+1}^2}{\epsilon'^2} \times \log \frac{15 \cdot 2^{a_{C+1}} \overline{\rho}_{C+1} \overline{\varphi}_{C+1} \tau_{C+1} \Lambda_{C+1} N}{\epsilon'}.
\end{aligned}$$

By replacing ϵ' by $\epsilon / \overline{\rho_{C+1}}$ and letting p_k take its new value $p_k + a_{C+1}$, we conclude that the statement holds for f^{C+1}. □

Now, the proof of (27) (and thus Theorem 5) stems from the relative deviation bounds of [16]. In particular, by Corollary 5 of [16], with probability at least $1 - \delta$,

$$V(x) \leq V_\gamma(x) \leq \hat{V}_\gamma(x) + 2\sqrt{\hat{V}_\gamma(x) \frac{\log \mathcal{N}(\frac{\gamma}{2}, \mathcal{F}_\gamma, 2N) + \log \frac{4}{\delta}}{N}} + 4 \frac{\log \mathcal{N}(\frac{\gamma}{2}, \mathcal{F}_\gamma, 2N) + \log \frac{4}{\delta}}{N},$$

where $\mathcal{F}_\gamma$ is the class of functions from $\mathcal{F}$ clipped at γ:

$$\mathcal{F}_\gamma = \{ f_{x,\gamma} \in [-\gamma, \gamma]^\Theta : f_{x,\gamma}(\theta) = \max\{-\gamma, \min\{\gamma, f_x(\theta)\}\}, \ f_x \in \mathcal{F} \}.$$

Since clipping $\mathcal{F}$ amounts to composing its functions with the 1-Lipschitz function $\max\{-\gamma, \min\{\gamma, \cdot\}\}$, Lemma 3 ensures that $\mathcal{N}(\frac{\gamma}{2}, \mathcal{F}_\gamma, 2N) \leq \mathcal{N}(\frac{\gamma}{2}, \mathcal{F}, 2N)$. Then, Lemma 6 applied with $\epsilon = \gamma/2$ gives the result (27).