A survey on scenario theory, complexity and compression-based learning and generalization

Roberto Rocchetta, Alexander Mey, and Frans A. Oliehoek

Abstract—**This work investigates formal generalization error bounds that apply to support vector machines in realizable and agnostic learning problems. We focus on recently observed parallels between results found in the statistical learning literature, like compression and complexity-based bounds, and novel error guarantees derived within scenario theory. Scenario theory provides non-asymptotic and distributional-free error bounds for models trained by solving data-driven decision-making problems. Relevant theorems and assumptions are reviewed and discussed. We propose a numerical comparison of the tightness and effectiveness of theoretical error bounds for support vector classifiers trained on several randomized experiments from thirteen real-life problems. This analysis allows for a fair comparison of different approaches from both conceptual and experimental standpoints. Based on the numerical results, we argue that the error guarantees derived from scenario theory are often tighter. This work promotes scenario theory as an alternative tool for model selection and generalization error analysis of support vector machines. In this way, we hope to bring the communities of scenario and statistical learning theory closer so that they can benefit from each other's insights.**

Index Terms—**Generalization theory, scenario optimization, PAC, compression, agnostic learning, support vector classifiers**

Nomenclature

$\mathcal{D}_m \subseteq \Delta^m$	data set of size m
$\delta = (x, y) \in \mathcal{D}_m$	sample (features and a label) in $\mathcal{D}_m$
$f \in \mathcal{F}$	function/model in a set of models
$\theta \in \Theta$	parameter vector in a set of parameters
$f^\star, \theta^\star$	optimized model/parameters
$\mathcal{A} : \mathcal{D}_m \to \mathcal{F}$	data-driven learning algorithm
$\mathfrak{C} : \Delta^m \to \Delta^{d_{cp}}$	a data compression rule
$y \in \mathcal{Y}$	label sample in an output set
$\mathcal{O}(\cdot), \Theta(\cdot), \Omega(\cdot)$	big-O notation [1]
$\lceil r \rceil$	smallest integer larger than $r \in \mathbb{R}$
$\lfloor r \rfloor$	largest integer smaller than $r \in \mathbb{R}$
$R(\theta) = R(f(\theta))$	true risk for model $f(\theta)$
$\hat{R}(\theta) = \hat{R}(f(\theta))$	empirical risk for $f(\theta)$
$V(\theta) = V(f(\theta))$	true margin violation probability for $f(\theta)$
$\hat{V}(\theta) = \hat{V}(f(\theta))$	empirical violation probability for $f(\theta)$
ϵ	threshold risk/robustness level
$\underline{\epsilon}, \overline{\epsilon}$	lower and upper error bounds
β	confidence level
$J(\theta), l(\theta)$	cost and loss functions
n_θ	number of model parameters
n_x	number of input features
ζ	slack variables
(w, b)	separating hyperplane parameters
$\mathcal{Z}$	Hilbert space
$\psi(x)$	kernel function
$\mathcal{S}^\star \subseteq \mathcal{D}_m$	a support set of minimum cardinality
$s^\star_m = \lvert \mathcal{S}^\star \rvert$	number of support constraints
d	VC dimensions (model complexity)
d_{cp}	size of the compression
γ_{CX}	tightness of complexity bounds
γ_{SB}	tightness of scenario bounds
γ_{CP}	tightness of compression bounds

I. Introduction

The generalization error, also known as risk or out-of-sample error, quantifies the ability of models to predict previously unseen data and plays a fundamental role in model selection for machine learning (ML) [2]. In practice, ML models are often chosen by minimizing an empirical estimate of their generalization error, for instance, applying k-fold cross-validation [3], [4], bootstrapping [5], jackknife [6], and leave-one-out methods [7]. These empirical approaches are well-established among practitioners and applied in diverse fields, including text classification [8] and categorization, [9], clustering [10], language processing [11], object [12] and fraud detection [13], unbalanced learning [14], pruning [10] as well as in distributed, federated, multitask and active learning [15]–[18]. However, empirical model selection methods can be computationally expensive, especially for complex models and large data sets, and the need to estimate the generalization error inevitably reduces the data available to train the models, which can be an issue under data scarcity. If a severe lack of training examples affects the study, for instance, when a data set is highly imbalanced or small, cross-validation and bootstrapping methods may result in unsatisfactory performance [19].

Unlike empirical model selection methods, *formal generalization error bounds* are mathematically derived and do not require test samples to estimate the generalization error. Instead, the generalization error can be bounded without using training data for empirical testing, speeding up the training and model selection. Over the years, significant research has focused on analyzing the theoretical properties of formal generalization error bounds and, recently, [20] proposed a data-free method for large-scale neural networks that only uses spectral proprieties of the weights to analyze the model

R. Rocchetta works at the University of Applied Sciences and Arts of Southern Switzerland, SUPSI-DACD-ISAAC, within the intelligent energy systems team, A. Mey is at the Department of Mathematics and Computer Science at Technical University of Eindhoven, Netherlands. F. A. Oliehoek works within the Technical University of Delft, Netherlands.

generalization. [21] introduced a Sharpness-Aware Learning Rate Scheduler to improve generalization by dynamically updating the learning rate of gradient-based optimizers, and [22] studied stability and convergence properties of generalization risk bounds for a particular kind of regularized distributed learning algorithms.

In this context, the probably approximately correct (PAC) learning framework [23] is one of the most widely applied to study the generalization error within the statistical learning literature, and prescribe formal generalization error bounds for ML models. In the PAC learning framework, a learner receives samples and must select a function (called the hypothesis) from a certain class of possible functions with good generalization properties. This framework has been used to compute guarantees for many models, inducing support vector models [24], graph kernels [25], majority voting classifiers [26], multi-view learners [27], and on domain adaptation and [28] multi-class domain adaptation problems [29], as well as on general classes of Boolean functions [30]. Several types of statistical learning bounds exist, and they can be based on the Bayes theorem in the PAC-Bayes framework [23], on an indicator of the model's complexity [31], or on the model's ability to compress the data [32]. Similarly to PAC-learning theory, scenario theory studies formal generalization error bounds for data-driven decision-making problems. Recently, error bounds introduced by [33] have been applied to anomaly detection [34], [35], interval regression [36], multi-agent learning [37], to construct predictive belief functions from data [38], and to study majority voting classifiers [39], and to robustly design controllers and other systems [40]–[42]. Based on the reviewed literature, only a few works used arguments from scenario theory to obtain generalization bounds for ML models like SVM, e.g., the works of S. Garatti and M. Campi [33], [43] that recently introduced formal scenario bounds for various models including SVM.

While attempts have been made to connect scenario theory and statistical learning theory [43], the literature is missing a clear comparison of the tightness of the bounds, especially from a pragmatic and quantitative/numerical standpoint. This work tries to fill this gap by proposing a numerical comparison of formal generalization error bounds from different theories. We focus on data-realizable and agnostic learning problems[1] and review and discuss underlying assumptions and theorems. In Figure 1 we summarize the main concepts we review in this paper. The comparison of the tightness of the bounds is proposed on randomized experiments from thirteen real-world data sets. For synthesis and clarity's sake, we focus our analysis on binary support vector machine (SVM) classifiers, both for soft-margin (agnostic) and hard-margin (data-realizable) cases. Note that, while the numerical analysis in this work focuses on SVM classifiers only, the reviewed theories apply to other ML models and to general classes of agnostic and realizable learning problems. Hence, one of the main contributions of this work is a fair comparison of different generalization theories and a numerical evaluation of the tightness of formal generalization error bounds.

The numerical results suggest that scenario bounds are often tighter, particularly when the generalization error is small, and better reflect the true risk for changing hyperparameters. With this, we hope to bring scenario theory forward to the Artificial intelligence (AI) community and connect it to known results from statistical learning theories such as the one based on the concepts of data compression and model complexity. Likewise, we hope that we present work done in statistical learning theory in an approachable manner for scientists working in scenario theory, so that they may take some inspiration for their work.

A. *Related literature*

This survey is motivated by recent theoretical results on the error of support vector models [33] and an expression of the research community on the need to investigate equivalences between different generalization error theories [44]–[46]. Various researchers have investigated error bounds from different theories to comprehend the relationship between the complexity of learning models and formal generalization guarantees achievable under different data availability scenarios. In [44], [46], the authors demonstrated the equivalence between PAC-learnability and compressibility, whilst [47] established a new connection between PAC-learning and stability theory. In [48], the authors focused on various online learning settings while [45] investigated an equivalence between PAC-learning and query-learning bounds. Kostas Margellos et al. [49] made the first attempt to connect compression learning with scenario theory and focused on bounding the out-of-sample error probability of realizable data-driven decision-making problems. Their results show that the issue of providing guarantees on the constraints violation probability reduces to a learning problem for an appropriately chosen algorithm that enjoys compression learning properties. Importantly, they show that ideas from scenario theory can strengthen or relax the consistency assumption to analyze learnability properties [49]. Licio Romao and collaborators [50] combined results from scenario theory and compression learning to derive tight error bounds for the solutions of realizable convex algorithms with discarded samples.

None of the reviewed works compared scenario bounds for agnostic learning with complexity-based and compression-based error bounds. Additionally, the literature lacks a numerical comparison of the tightness of the formal error bounds under different data availability conditions. Therefore, further research is needed to link scenario, complexity, and compression-learning theory for realizable and agnostic learning settings, to identify the strengths and weaknesses of different approaches, and to popularize theoretical error

[1] See Section II-C for a definition of realizable and agnostic learning problems.

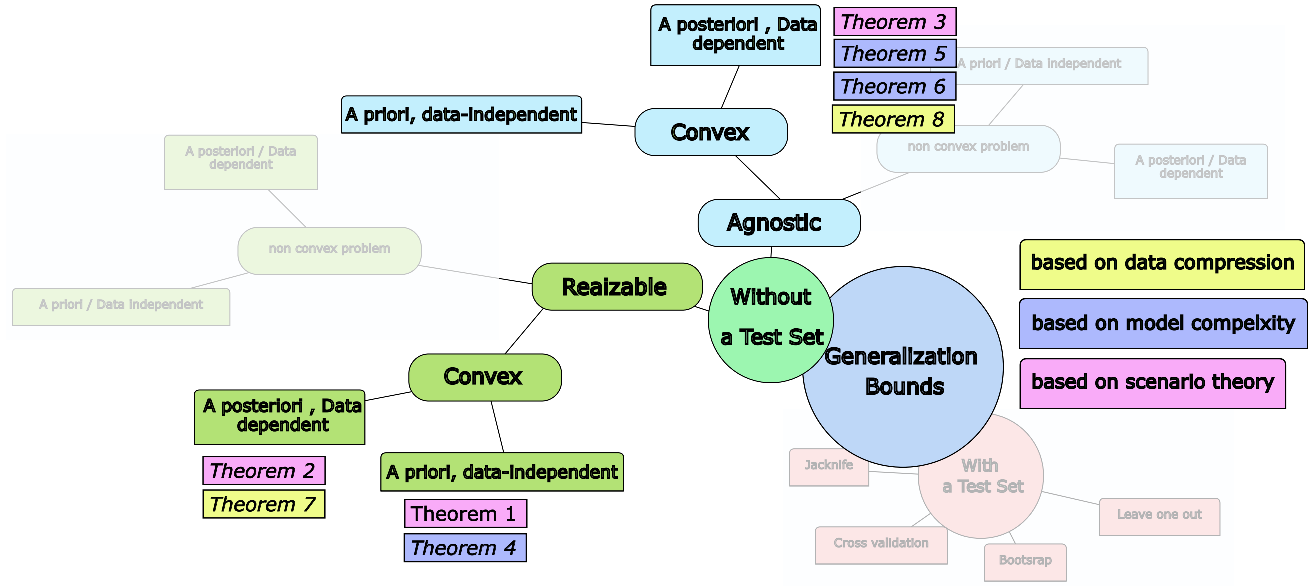
A posteriori , Data dependent
Theorem 3
Theorem 5
Theorem 6
Theorem 8
A priori / Data independent
A priori, data-independent
Convex
non convex problem
A posteriori / Data dependent
A posteriori / Data dependent
Agnostic
non convex problem
A priori / Data independent
based on data compression
Realizable
Without
a Test Set
based on model compelxity
Generalization
Bounds
based on scenario theory
Convex
A posteriori , Data dependent
Theorem 2
Theorem 7
A priori, data-independent
Theorem 1
Theorem 4
Jacknife
With
a Test Set
Leave one out
Cross validation
Bootsrap

Fig. 1. A scheme summarizing the reviewed generalization error bounds and theorems. Note that our paper focuses on formal error bounds that do not require empirical estimation of the risk from a test set and primarily focused on scenario bounds both for agnostic and reliable settings.

bounds among practitioners. This overview demonstrates the need for further research in the field and provides additional evidence supporting the relevance of this work. The rest of this paper is organized as follows: Section II presents the mathematical background, and section III reviews theories for prescribing generalization error bounds without a test set. A numerical evaluation of the tightness of the bounds on thirteen data sets is carried out in Section V. Section VI closes the paper with a discussion of the results, remarks and conclusions.

II. Preliminaries

Let us consider a data set $\mathcal{D}_m = \{\delta_i\}_{i=1}^m$ with m independent and identical distributed (iid) samples drawn from an unknown probability space $(\Delta, \mathfrak{F}, \mathbb{P})$, comprising an event space Δ, equipped with a σ-algebra $\mathfrak{F}$, and a fixed probability measure $\mathbb{P} : \mathfrak{F} \to [0,1]$. The probability $\mathbb{P}$ is assumed unavailable, which is generally the case in practice[2]. A sample $\delta \in \mathcal{D}_m$ will be also called a *scenario*.

A. Supervised learning

A scenario $\delta = (x, y) \in \mathcal{D}_m$ contains a vector of explanatory variables $x \in \mathcal{X} \subseteq \mathbb{R}^{n_x}$ and target variables $y \in \mathcal{Y}$. Based on the samples in $\mathcal{D}_m$ we have to choose a function $f \in \mathcal{F}$, with $f : \mathcal{X} \to \mathcal{Y}$, from a function class $\mathcal{F}$ with the goal that f is a good predictor of y, given a newly seen sample x. The process of learning a predictive model (a function), can be defined as a generic data-driven decision-making problem,

$$\mathcal{A} : \Delta^m \to \mathcal{F}, \ \ m = 0, 1, 2, ..., \tag{1}$$

where $\mathcal{A}$ is a map from the samples space $\Delta^m = \Delta \times \Delta \times ...$ (m times) and the decision space $\mathcal{F}$. Note that the data set $\mathcal{D}_m$ is a random realization from the event space Δ^m and, without loss of generality, the map $\mathcal{A}$ can be seen as a sophisticated optimization method or a simple heuristic to select a function $f \in \mathcal{F}$ based on the available data $\mathcal{D}_m \in \Delta^m$. For instance, the decision-making problem $\mathcal{A}$ for selecting a predictive model is generally focused on the identification of a function $f^\star := \mathcal{A}(\mathcal{D}_m) = \arg\min_{f \in \mathcal{F}} \sum_{\delta \in \mathcal{D}_m} l_f(\delta)$ that achieves a small probability of prediction error, i.e., that minimizes the expectation of a loss function $l_f(\delta)$.

B. Binary Classification

Binary classification is a special type of supervised learning where new observations of explanatory variables x must be categorized into one of two classes. For binary classification problems the dimension of δ is, therefore, $n_x + 1$ as $y \in \{-1, +1\} \subset \mathcal{Y}$ and an indicator function for the loss can be considered as follows:

$$l_f(x, y) = \begin{cases} 1, & \text{if } y \neq f(x) \\ 0, & \text{otherwise} \end{cases}$$

where, if $y \neq f(x)$ the loss function results in $l_f(x, y) = 1$ and the model f fails to classify the class of x correctly. The expected value of this loss function is the misclassification probability:

$$R(f) = \mathbb{P}[\delta \in \Delta : l_f(\delta) = 1],$$

where R is also known as error probability or risk. As we assume to not have direct access to $\mathbb{P}$ we cannot evaluate $R(f)$ for a given f precisely and, thus, we have to find other means of choosing a suitable candidate model f. For instance, minimizing an empirical (samples-based) estimate of the error probability given by:

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} l_f(\delta_i) \tag{2}$$

[2] Here, albeit unavailable, the measure $\mathbb{P}$ is assumed stationary (fixed).

Here n is the number of available samples for the estimation (test samples) and $l_f(\delta_i)$ is the indicator function for the misclassification event $f(x_i) \neq y_i$.

C. Realizable and agnostic problems

Given a probability measure $\mathbb{P}$ and a hypothesis class $\mathcal{F}$ we call the learning problem *realizable* if and only if there exists a function $f^* \in \mathcal{F}$ such that $R(f^*) = 0$. If such an f^* does not exist we call the learning problem *agnostic*. A further distinction can be made between learning problems that are realizable for probability distributions of δ and learning problems that are realizable for a specific data set $\mathcal{D}_m$, [44]. We will refer to the latter problems as *data-realizable* problems. In contrast, a problem is called *data-agnostic* if it is not data-realizable for $\mathcal{D}_m$. For instance, consider the following binary classification problem:

$$\begin{aligned} f^\star = &\arg\min_{f \in \mathcal{F}} J(f) \\ \text{s.t.: } &l_f(\delta_i) = 0 \ \ \forall \delta_i \in \mathcal{D}_m. \end{aligned} \tag{3}$$

where $J(f)$ is a cost function and the samples in $\mathcal{D}_m$ define m constraints on the miss-classification errors. This simple learning problem is data-realizable for $\mathcal{D}_m$ if and only if a function $f^\star$ exists such that $l_{f^\star}(\delta_i) = 0$ for all $\delta_i \in \mathcal{D}_m$. In other words, if the problem admits a non-empty feasible set and at least one solution exists the satisfies all the constraints on the loss function. However, feasibility does not guarantee the problem is realizable for all unobserved realizations of $\delta \in \Delta$ from the same distribution. Our work will focus on both data-realizable and agnostic problems.

D. Support vector machines

Support vector machines [51] have been historically developed as binary non-probabilistic classifiers and are nowadays one of the most widely applied ML models in practice. An SVM classifier is a binary classification function $f : \mathcal{X} \to \{-1, +1\}$ that maximizes the margin between the decision boundary of f and the two classes defined by the labels in $\mathcal{D}_m$. In the following paragraphs, we introduce the standard notation and definitions of SVMs.

1) Linear hard-margin SVM: A linear hard-margin SVM program is defined as follows:

$$(w^\star, b^\star) = \arg\min_{\substack{b \in \mathbb{R} \\ w \in \mathbb{R}^{n_x}}} \|w\| \tag{4}$$

$$\text{s.t.: } \ y_i(w \cdot x_i - b) \geq 1, \ i = 1, ..., m \tag{5}$$

where $w \in \mathbb{R}^{n_x}$ is the normal vector to the hyperplane defining the boundary of the two classes, b is the bias term, the parameter $\frac{b}{||w||}$ determines the offset of the hyperplane and $|| \cdot ||$ is the L_2 norm operator. In view of the earlier notation, our model class $\mathcal{F}$ is now given by $\mathcal{F}_{\text{lin}} := \{x \to \text{sign}(w \cdot x + b) \mid w \in \mathbb{R}^{n_x}, b \in \mathbb{R}\}$. Program (4) maximizes the width of a hyperplane, proportional to $\frac{1}{||w||}$, separating the space $\mathcal{X}$ in two regions, one dedicated to each class. The term hard-margin means that the constraints $y_i(w \cdot x_i - b) \geq 1, i = 1, ..., m$ must be satisfied. If the learning problem is realizable with respect to a linear hypothesis class, the optimal hyperplane defined by the parameters $(w^\star, b^\star)$ always exists.

2) Non-linear soft-margin SVM: Two standard extensions of the linear hard-margin SVM are towards a non-linear function class and towards a program that relaxes the separability constraints. The non-linearity is achieved by introducing a kernel function $\psi(x) : \mathcal{X} \to \mathcal{Z}$, mapping the physical space $\mathcal{X}$ to a Hilbert space $\mathcal{Z} \subseteq \mathbb{R}^{n_z}$, whilst margin relaxation is achieved introducing slack variables ζ in the constraints of program (4). Note that the Hilbert space is generally of higher dimension and the function $\psi(x)$ does not need to be explicitly defined [33]. All relevant computations rely only on the evaluation of a kernel $K(x_k, x_j) := \psi(x_k) \cdot \psi(x_j)$, i.e., the evaluation of inner products. The kernel $K(\cdot, \cdot)$ can be used to operate in $\mathcal{Z}$ without the need for actually computing the coordinates of the measurements in the Hilbert space in an explicit way [43], [52]. The combination of kernel and constraint relaxation gives rise to a non-linear soft-margin SVM program:

$$\begin{aligned} (w^\star, b^\star, \zeta^\star) = &\arg\min_{\substack{b \in \mathbb{R}, w \in \mathbb{R}^{n_z} \\ \zeta \in \mathbb{R}^m_+}} \|w\| + \rho \sum_{i=1}^{m} \zeta_i : \\ \text{s.t:} \quad &y_i(w \cdot \psi(x_i) - b) \geq 1 - \zeta_i, \ i = 1, ..., m \end{aligned} \tag{6}$$

where ζ is a vector of m non-negative slack variables and $\rho > 0$ is a scalar regularization parameter weighting the cost of margin violations. Program (6) seeks an optimal hyper-plane $(w^\star, b^\star)$ which linearly separates the data in the space $\mathcal{Z}$ and minimizes the cost of margin violations given by $\zeta^\star$. A linear separator in the Hilbert space will map back to a non-linear separator in $\mathcal{X}$. Hence, the hypothesis class $\mathcal{F}$ is now given by $\mathcal{F}_{nlin} := \{x \to \text{sign}(w \cdot \psi(x) - b) \mid w \in \mathbb{R}^{n_z}, b \in \mathbb{R}\}$ and the optimized hyperparameters $\theta^\star = (w^\star, b^\star)$ defines a unique classifier $f^\star = f(x; \theta^\star) \in \mathcal{F}_{\text{nlin}}$. Note that although we use the kernel function ψ explicitly for ease of notation, all those computations may also be done using only evaluations of the kernel K, see for example Equation (19) in [52].

3) Margin violation and misclassification error: Label predictions can be assigned to new observations x by $\hat{y} = f(x; \theta^\star)$ and misclassification occurs if $y \neq \hat{y}$ or, equivalently, if $y(w^\star \cdot \psi(x) - b^\star) < 0$. The misclassification probability (error/risk) for an SVM classifier is defined by:

$$R(\theta^\star) := \mathbb{P}[y(w^\star \cdot \psi(x) - b^\star) < 0]. \tag{7}$$

The margin violation event occurs if a new sample pair (x, y) violates the constraint in the training program and the probability of margin violation $V(\theta^\star)$ is defined as follows:

$$V(\theta^\star) := \mathbb{P}[y(w^\star \cdot \psi(x) - b^\star) < 1]. \tag{8}$$

Note that, by definition, the margin violation probability bounds from above the misclassification probability, i.e., $V(\theta^\star) \geq R(\theta^\star)$.

4) Other SVM models: Many variants of the hard-margin and soft-margin SVM models have been proposed in the literature to tackle regression tasks [53], [54], fault and anomaly detection [35] for prognostics of industrial assets and components [55]–[57]. For instance, support vector data descriptor (SVDD) [58] and one-class SVM models [59] are often used for anomaly detection, where SVDD identify abnormalities beyond an optimized spherical region in the kernel space and one-class SVM beyond an optimized hyperplane. Both models undergo training through convex optimization constrained by random scenarios, making the formal bounds discussed herein relevant. Similarly, twin support vector machines (TWSVM) undergo convex optimization of non-parallel hyperplanes for class separation, often yielding improved accuracy and generalization compared to standard SVMs. TWSVM were originally introduced for pattern classification [60], [61] and later extended to tackle active learning tasks [62], multi-class learning, [63], and other applications [64], [65]. A comprehensive TWSVM survey [66] covered clustering [67], semi-supervised classification [68], and outlier detection [69]. Although the theoretical approaches reviewed in this work are applicable to advanced SVM models and general classes of learning problems, including neural network models [70], [71], to ease the presentation, this work only focuses on formal generalization for traditional binary SVM classifiers.

E. Generalization error bound

As we only have a finite amount of data available, the values of R and V are unknown. Nonetheless, generalization error bound analysis can be used to find upper and lower bounds on these probabilities. Formally, given an algorithm $\mathcal{A}$ as defined above, a generalization error bound is a function $B_{\mathcal{A}}(\beta, m, I)$ of a confidence parameter $0 < \beta < 1$, the sample size m, and other (potentially algorithm dependent) parameters I such that, with probability of at least $1 - \beta$ over the random sample, it holds that

$$R(\mathcal{A}(\mathcal{D}_m)) \leq B(m, \beta, I). \tag{9}$$

In the next section, we present an overview of some of the most popular theories for those bounds and we then focus more in detail on complexity-based bounds, compression learning bounds and their link to new bounds obtained via scenario theory. Then, in the following section, specific results are presented and later used as a basis for the numerical comparison.

III. Generalization error bounds: An Overview

Several types of formal generalization error bounds can be found in the scenario theory and statistical learning theory literature and can be summarized as follows:

PAC-Bayes methods use, amongst other concepts, a prior distribution over function class $\mathcal{F}$ to find generalization error bounds. The PAC-Bayesian theory has been successfully used in a variety of topics, including sequential learning, classification, and analysis of heavy-tailed data, e.g., [28], [72], [73], for ranking and bounding probabilities of non-iid samples, [74], and for enhancing the generalization of regularized Neural Networks [75]. The PAC-Bayes framework requires a definition of a prior distribution over the different classifiers (generally before observing the data), and the definition of a suitable likelihood that is needed to approximate a posterior distribution (once the data is collected).

Complexity-based methods quantify the generalization of ML models based on a measure of the capacity (complexity, expressive power, or richness) of the class of models $\mathcal{F}$. If the models in $\mathcal{F}$ are complex, the optimized $f^\star$ is more likely to over-fit the data and generalize poorly. The Vapnik–Chervonenkis (VC) dimension, as originally introduced by Vapnik [31], is a complexity measure which helps to bound the difference of the empirical risk and true risk uniformly over the model class. Given a hypothesis class $\mathcal{F}$, the VC-dimension is the maximum number of features x that can be labelled in all possible ways with functions from $\mathcal{F}$. Other examples of complexity-based bounds include extensions based on the VC-dimension [76], FAT-shattering dimension, covering number, global and local Rademacher complexity [77], union bound and shell bound methods [78]. In the case of learning a hyperplane, which is the essence of an SVM, one can relate the VC dimension to the margin between the hyperplane and the sample $\mathcal{D}_m$. Complexity-based bounds apply to deterministic classification rules and algorithms but are not usable for learning algorithms for which $\mathcal{F}$ is unknown or data-dependent.

Compression-based methods estimate algorithms' generalization in terms of their ability to create a reduced-size representation of the data, i.e., the algorithms' ability to compress samples [79]. If the data representation is small (relative to the sample size), the compression rate will be high, and the algorithm will likely generalize well. In contrast to complexity-based bounds, which provide bounds based on a capacity measure of $\mathcal{F}$, compression-based error bounds are data-dependent, as they depend on both the particular choice of learning algorithms $\mathcal{A}$ and the random data in $\mathcal{D}_m$. Refer to [80], [81] for other examples of data-dependent error bounds. Littlestone and Warmuth [32] originally introduced the concept of compression for zero/one loss functions, relating the bounds to early works on Kolmogorov complexity. A compression function maps the data to a subset of the original data (a compression set) that suffice to reconstruct the resulting model $f^\star$. The compression function then describes how much we may compress the data while ensuring that $\mathcal{A}$ provides the same $f^\star$ under the uncompressed $\mathcal{D}_m$. The main limitation of compression methods are the huge compression rates needed to obtain informative bounds, rates which are not always achievable in practice. A conceptually close approach is the coreset learning theory [82], which relates to the optimal subsampling literature and defines generalization as a small weighted subset set of samples that approximates this loss for every element in a query set. Compression bounds are already available for complex models, including deep neural networks [83], leading to orders of magnitude better performance when compared to PAC-Bayes and complexity-based bounds.

Scenario theory was, differently from previous approaches, originally introduced to investigate the error probability of convex optimization problems with randomized constraints [84], [85]. Those are learning problems where samples of uncertain factors define deterministic convex constraints [86], and the hypothesis class defines a parametric family of convex functions. Scenario theory has been extensively studied for hard-constrained (data-realizable) convex programs [86]–[89] and recently extended to non-convex cases [90] and more abstract classes of decision-making algorithms [91]. Scenario bounds can be derived before solving an optimization problem or can be tailored to the optimized solutions. The former bounds are known as *a-priori* and, like complexity-based bounds, are derived independently from the data. For instance, one of the most acclaimed results from scenario theory proved that the error distributions of solutions to convex problems are bounded by beta distributions whose parameters can be a-priori determined by the number of training samples and the number of optimization parameters. This bound was proven to be tight (exact) for a whole class of problems named fully-supported [86]. Differently, scenario bounds tailored to the specific models are known as *a-posteriori* bounds and are based on a data-dependent hypothesis, like compression-based bounds. For instance, empirical error levels of the solution of min-max problems have been shown to follow a Dirichlet distribution, whose marginals are beta distribution [88]. Scenario theory combines ideas from complexity-based methods and compression-based methods. Similarly to VC-bounds, the complexity of the class $\mathcal{F}$ defines data-independent (a-priori) scenario bounds. Like compression-based methods, a-posteriori scenario bounds depend on a notion of how much one may compress the given learning rule $\mathcal{A}$ for a given $\mathcal{D}_m$. A learning rule $\mathcal{A}$ having a higher degree of compression leads to better generalization guarantees, thus, an interesting link is established between the complexity of the candidate class $\mathcal{F}$ and the joint complexity of a specific $f^\star \in \mathcal{F}$ and the observed data $\mathcal{D}_m$.

IV. Generalization error bounds: Specific Results

We now present specific generalization error results, first from scenario theory and then from statistical learning theory.

A. Scenario Theory

In scenario theory, a scenario program $\mathcal{A}(\mathcal{D}_m)$ defines a general class of data-driven decision-making problems, as in equation (1). Without loss of generality and to ease the presentation, we focus on a specific class of programs where a set of deterministic constraints are defined by the random samples in $\mathcal{D}_m$, like in SVM training programs. In the next sections, we formally introduce convex hard-constrained (data-realizable) and soft-constrained (agnostic) scenario programs and establish their link to SVM training programs (4) and (6).

*1) **Hard-constrained scenario program**:* A hard-constrained convex scenario optimization program is defined as follows:

$$\min_{\theta\in\Theta} J(\theta), \text{ s.t. } f(\theta,\delta_i) \le 0,\ \delta_i \in \mathcal{D}_m \tag{10}$$

where $\theta \in \Theta \subseteq \mathbb{R}^{n_\theta}$ is a vector of design variables constrained in a closed convex set Θ, $J : \Theta \mapsto \mathbb{R}$ is a convex cost function and $f(\theta,\delta) : \Theta \times \Delta \to \mathbb{R}$ is a convex function in θ defining m hard-constraints in (10). An optimal feasible design $\theta^\star$ must satisfy $f(\theta^\star,\delta_i) \le 0$ for all $i = 1, ..., m$, with no exception and this often leads to a feasibility issue.

*2) **Soft-constrained scenario program**:* A soft-constrained reformulation of (10) is given by:

$$\min_{\substack{\theta\in\Theta \\ \zeta\in\mathbb{R}^m_+}} J(\theta) + \rho\sum_{i=1}^{m} \zeta_i : \tag{11}$$
$$\text{s.t. } f(\theta,\delta_i) \le \zeta_i, \delta_i \in \mathcal{D}_m$$

where ζ is a m-dimensional vector of non-negative slack variables. A $\zeta_i = 0$ means that the hard-constraint imposed by the i^{th} sample is satisfied, i.e., $f(\theta,\delta_i) \le 0$. On the other hand, a $\zeta_i > 0$ implies a violation of the hard constraint. Note that SVM programs are special classes of $\mathcal{A}(\mathcal{D}_m)$ where the cost and constraint functions are defined by

$$J(\theta) = ||w|| \qquad f(\theta,\delta) = 1 - y_i(w \cdot \psi(x_i) - b), \tag{12}$$

with $\delta = (x,y)$ being a sample and $\theta = (w,b)$ the parameters of the separating hyperplane. When a kernel operator is applied, the number of design variables becomes $n_\theta = n_z + 1$, i.e. the dimension of the Hilbert space plus one due to the bias term.

*3) **Assumptions and definitions**:* Scenario theory can be used to assess how well an optimal design θ^*, so a solution of the optimization programs (10) or (11), generalizes to yet unseen situations $\delta \in \Delta$. Definitions, assumptions and relevant theorems needed to compute generalization bounds will be presented next.

Definition 1: (**Violation probability**) The probability

$$V(\theta^\star) = \mathbb{P}\left[\delta \in \Delta : f(\theta^\star,\delta) > 0\right] \tag{13}$$

is called violation probability. Given a reliability parameter $\epsilon \in [0,1]$, a design $\theta^\star$ is called ϵ-robust if $V(\theta^\star) \le \epsilon$. Note that for SVM programs, the violation probability coincides with the 'true' margin violation probability (see the Appendix for the definition) and, thus, an ϵ-robust SVM $\theta^\star$ satisfies $R(\theta^\star) \le V(\theta^\star) \le \epsilon$, i.e., a bound on the worst-case classification error probability.

Definition 2: (**Non-reducible support set**) A support set $\mathcal{S} \subseteq \mathcal{D}_m$ is a k-tuple $\mathcal{S} = \{\delta_{i_1}, ..., \delta_{i_k}\}$ for which the solutions of the scenario program $\mathcal{A}(\mathcal{S})$ and program $\mathcal{A}(\mathcal{D}_m)$ are identical. A set $\mathcal{S}^\star \subseteq \mathcal{S}$ is non-reducible if for any $\delta \in \mathcal{S}^\star$ the solution of $\mathcal{A}(\mathcal{S}^\star \setminus \delta)$ differs from the one of $\mathcal{A}(\mathcal{D}_m)$, i.e., the support set is of minimal cardinality. A scenario program generally admits several support sets and the set $\mathcal{S}^\star$ with the smallest cardinality renders the best bounds. The dimension of $\mathcal{S}^\star$ will be denoted as $s^\star_m = |\mathcal{S}^\star|$, where $|\cdot|$ is the cardinality operator.

Assumption 1: *(Existence and uniqueness) For every data sequence $\mathcal{D}_m$, the design solution $\theta^\star$ of $\mathcal{A}(\mathcal{D}_m)$ exists and is unique.*

Assumption 2: *(Non-degeneracy) For any positive integer $m \in \mathbb{N}_0$ and scenario set $\mathcal{D}_m$, the solution of the scenario program $\mathcal{A}(\mathcal{D}_m)$ coincides with probability 1 with the solution of $\mathcal{A}(\mathcal{S}^\star)$.*

When program (10) is convex, non-degeneracy is a mild assumption since support constraints in $\mathcal{S}^\star$ are always active, and a (possibly reducible) support set can be easily identified [92]). In fact, the solution of convex learning programs is generally unaffected by removing inactive scenarios for which $f(\theta^\star, \delta) < 0$ at the optimum. In the general non-convex case, however, $\mathcal{S}^\star$ might include non-active constraints and the non-degeneracy assumption rarely holds. Hence, the removal of a single non-active constraint, i.e., the removal of samples δ for which $f(\theta^\star, \delta) < 0$, can yield a new optimum having a smaller cost [92]. A recent extension of the theory allows relaxing the non-degeneracy assumption [93] and allows extending the scope of scenario theory to a broader domain of learning problems where the sample constraints are non-convex functions of the model parameters and observations.

*4) **Bounds for hard-constrained programs**:*

Theorem 1: [86, Theorem 1] Under assumptions 1, 2, stationary $\mathbb{P}$ and iid samples in $\mathcal{D}_m$ the distribution of $V(\theta^\star)$, for $\theta^\star$ being the solution of (10), is bounded by a Beta distribution:

$$\mathbb{P}^m[V(\theta^\star) > \epsilon] \leq \sum_{k=0}^{n_\theta - 1} \binom{m}{k} \epsilon^k (1-\epsilon)^{m-k} = \beta \qquad (14)$$

where n_θ is the number of design variables, $\beta \in [0, 1]$ is a confidence parameter, and $\mathbb{P}^m$ is a product probability due to independence of the m samples.

Theorem 1 can be easily derived from Helly's Theorem, showing that $s_m^\star \leq n_\theta$ for convex programs under the given assumptions. Note that the bound ϵ is data-independent as it can be a-priori computed, i.e., it is obtained before solving (10), and it only requires the number of design variables n_θ, number of samples m, and a desired confidence β. If program (10) is fully supported, that is, if $s_m^\star = n_\theta$ with probability 1, Eq. (14) holds with the equality sign. An extension of Theorem 1 allows for k samples δ to be intentionally removed from the data set $\mathcal{D}_m$, for instance, the ones making program (10) unfeasible. This approach makes (10) a data-realizable problem. As a result, the optimized $\theta^\star$ enjoys an improved $J(\theta^\star)$ for the cost of a weaker certificate of generalization [86]. However, many real-life problems are only partially supported ($s_m^\star < n_\theta$) and the following Theorem 2 renders tighter bounds.

Theorem 2: [94, Theorem 2] Consider a convex scenario program defined as in (10). Under assumptions 1, 2, stationary $\mathbb{P}$ and iid samples in $\mathcal{D}_m$ the solution $\theta^\star$ of (10) satisfies

$$\mathbb{P}^m[V(\theta^\star) > \epsilon(s_m^\star)] \leq \beta, \qquad (15)$$

where the reliability $\epsilon(k) = 1 - t(k)$ is the unique solution in [0,1] of the following polynomial equation in t for any $k \in \{0, ..., n_\theta\}$:

$$\frac{\beta}{m+1} \sum_{j=k}^{m} \mathfrak{B}_j(t; k) - \binom{m}{k} t^{m-k} = 0 \qquad (16)$$

Here $\mathfrak{B}_j(t; k) = \binom{j}{k} t^{j-k}$ is a binomial expansion.

Theorem 2 gives a generalization bound $V(\theta^\star) \leq \epsilon(s_m^\star)$ at a confidence level $1 - \beta$ and reliability parameter $\epsilon(s_m^\star)$ determined from (16). In contrast to (14), $\epsilon(s_m^\star)$ is a-posteriori computed by enumerating support scenarios $s_m^\star$ in correspondence of $\theta^\star$.

*5) **Bounds for soft-constrained programs**:* To extend the scope of scenario-based generalization bounds to soft-constrained problems, like the one in equation (11), a technical assumption of *non-accumulation* is required. The assumption states that, for every $\theta \in \Theta$ and $a \in \mathcal{R}$, the function $f(\theta, \delta)$ does not have concentrated mass, i.e., $\mathbb{P}[\delta : f(\theta, \delta) = a] = 0$. This assumption is generally satisfied when δ admits a probability density function.

Theorem 3: [91, Theorem 4] Consider a convex scenario program as in equation (11). Given the aforementioned assumptions, stationary $\mathbb{P}$ and iid samples in $\mathcal{D}_m$ the probability $V(\theta^\star)$ is bounded by:

$$\mathbb{P}^m[\underline{\epsilon}(s_m^\star) \leq V(\theta^\star) \leq \overline{\epsilon}(s_m^\star)] \geq 1 - \beta, \qquad (17)$$

where $\underline{\epsilon}(k) = \max\{0, 1 - \overline{t}(k)\}$, $\overline{\epsilon}(k) = 1 - \underline{t}(k)$ and $\{\underline{t}(k), \overline{t}(k)\}$ are solutions of a polynomial equation in t:

$$\mathfrak{B}_m(t; k) = \frac{\beta}{2m} \sum_{j=k}^{m-1} \mathfrak{B}_j(t; k) + \frac{\beta}{6m} \sum_{j=m+1}^{4m} \mathfrak{B}_j(t; k) \qquad (18)$$

where $k \in \{1, ..., m-1\}$ is the number of support constraints. For the special case $k = m$, the upper bound is set to $\overline{\epsilon}(k) = 1$ and the lower bound is obtained solving

$$1 = \frac{\beta}{6m} \sum_{j=m+1}^{4m} \mathcal{B}_j(t; k). \qquad (19)$$

For a soft-constrained SVM design, Theorem 3 gives high probability upper and lower bounds on the probability of margin violations. Here $s_m^\star$ is the number of support vectors, i.e. the number of samples for which $1 - y(w^\star \psi(x) - b^\star) \geq 0$. If $w^\star = 0$, $s_m^\star$ is the number of data points whose label belongs to the class with fewer elements [33].

We refer to the bounds introduced in this section as the *scenario bounds*. We introduce generalization bounds from the statistical learning literature, based on complexity and compression, next. For our comparison, we use the most recent results on generalization bounds for SVMs.

B. Complexity-based bounds for SVM

For realizable cases, [24] solved a longstanding problem by showing that the error of an SVM drops as $\Theta(\frac{d}{m}+\log(\frac{1}{\beta}))$, where d denotes the VC-dimension. The specific bound they provide is defined as follows:

Theorem 4: [24, Theorem 15] Under the realizability assumption from Section II-C, it holds that for all data set sizes $m > 2(d+1)$, with probability of at least $1-\beta$,

$$R(\theta^\star) \leq \frac{2}{m-2(d)}\left((d)\ln(4)+\ln\left(\frac{1}{\beta}\right)\right), \tag{20}$$

where $\theta^\star = (w^\star, b^\star)$ is the solution of (4) and d again denotes the VC-dimension of the SVM and is equal to $d = n_x + 1$.

Next, we present two results in the agnostic case, taken from [95]. The first theorem bounds the generalization error in terms of the margin of the SVM. The second bound is more general and not directly targeted at SVMs, as it uses the VC dimension for general linear predictors $d = n_x + 1$. In view of our earlier discussions, the margin-based bound is an a-posteriori bound, as it needs to know the margin of the resulting SVM. The VC dimension-based bound is an a-priori bound, as the VC dimension is known before any learning.

Theorem 5: [95, Theorem 15] Let $\theta^\star = (w^\star, b^\star)$ be the solution of the SVM program (6) and let $l^\star_{SVM}(x,y) = \min\{\max\{0, 1-y(w^\star\cdot x+b^\star)\}, 1\}$ be the so called ramp loss. If $r = \max\limits_{1\leq i\leq m}\|x_i\|$, then with probability of at least $1-\beta$ it holds that:

$$R(\theta^\star) \leq \frac{1}{m}\sum_{i=1}^{m} l^\star_{SVM}(x_i,y_i) +$$
$$\frac{1}{\sqrt{m}}\left(4\sqrt{(\Lambda+1)^2}+3\sqrt{\ln(\frac{\pi^4\Lambda^2}{18\beta})}\right). \tag{21}$$

where $\Lambda = \lceil r\rceil\lceil\|w^\star\|\rceil \in \mathbb{N}$ is the product of smallest integers larger than r and $\|w^\star\|$.

The second term on the right-hand side of Equation 21 is derived from Rademacher complexity theory, whilst the first term provides a lower bound on the empirical risk estimated from the slack variables in (6), i.e., $\frac{1}{m}\sum_{i=1}^{m} l^\star_{SVM}(x_i,y_i) \leq \frac{1}{m}\sum_{i=1}^{m}\zeta^\star_i$.

Theorem 6: [77, Corollary 3.19] For all predictors $f(\theta) \in \mathcal{F}_{\text{lin}}$, so also for the SVM solution $f(\theta^\star)$, it holds that

$$R(\theta) \leq \hat{R}(\theta) + \tag{22}$$
$$\frac{1}{\sqrt{2m}}\left(\sqrt{4(d+1)\ln(\frac{em}{d})}+\sqrt{\ln(\frac{1}{\beta})}\right), \tag{23}$$

where $d = n_x + 1$ is the VC dimension of $\mathcal{F}_{\text{lin}}$ and e is Euler's constant.

Note that equation (22) has in comparison to (21) an additional logarithmic term $\ln(m)$ in the number of samples, which can in principle be removed.

C. Compression-based bounds for SVM

Similarly to a-posteriori scenario bounds, compression bounds depend on the number of data points that are strictly necessary to reconstruct the optimum. This quantity is known as compression size in compression learning and is equivalent to the support set size $|\mathcal{S}|$ in scenario theory. Note that $\mathcal{S}$ does not have to be of minimal cardinality for the bounds to apply. However, smaller values of the compression rates led to better generalization error guarantees.

In order to formally introduce compression learning bounds, we will make use of the mathematical framework presented in [44], [46]. We consider a learning algorithm $\mathcal{A}$ and a compression rule, $\mathfrak{C}$, where we call $\mathcal{A}$ permutation invariant if the mapping does not depend on the ordering of the input data set $\mathcal{D}_m$. The rule $\mathfrak{C} : \Delta^m \rightarrow \Delta^{d_{cp}}$ compresses the data $\mathcal{D}_m$ to a smaller data set $\mathcal{D}_{d_{cp}} \subset \mathcal{D}_m$. A compression scheme can be any rule that identifies a compression set $\mathcal{D}_{d_{cp}}$ such that the SVM classifier obtained from the original set $\mathcal{D}_m$ is exactly the same as the SVM obtained from the compressed set $\mathcal{D}_{d_{cp}}$. Clearly, the compression set coincides with the set of support scenarios $\mathcal{S}$ as defined in scenario theory, which is equivalent to the number of support vectors in the case of SVMs, showing a clear parallel between the two approaches. As in the previous settings we first present a result in the realizable and then in the agnostic setting.

Theorem 7: [96, Theorem 1] Consider a realizable learning problem and let $\mathcal{A} : \Delta^m \rightarrow \mathcal{F}$ be a permutation invariant learning rule with $\hat{R}[\mathcal{A}] = 0$. Let $d_{cp}(\mathcal{D}_m)$ be the size of a compression scheme for $\mathcal{D}_m$, such that $\theta^\star = \mathcal{A}(\mathcal{D}_m) = \mathcal{A}(\mathcal{D}_{d_{cp}})$. For any $\mathbb{P}$, $m \in \mathbb{N}$ with probability at least $1-\beta \in [0,1]$ over the sampling of $\mathcal{D}_m$ it holds that

$$R(\theta^\star) \leq$$
$$\frac{1}{m-d_{cp}}\left(\ln\binom{m}{d_{cp}}+\ln(m)+\ln\left(\frac{1}{\beta}\right)\right). \tag{24}$$

Theorem 8: [96, Theorem 2] Consider an agnostic learning problem and let $\mathcal{A} : \Delta^m \rightarrow \mathcal{F}$ be a permutation invariant learning rule. Let $d_{cp}(\mathcal{D}_m)$ be the size of a compression scheme for $\mathcal{D}_m$, such that $\theta^\star = \mathcal{A}(\mathcal{D}_m) = \mathcal{A}(\mathcal{D}_{d_{cp}})$. For any $\mathbb{P}$, $m \in \mathbb{N}$, training data set $\mathcal{D}_m$, and $\beta \in [0,1]$, with probability at least $1-\beta$ it holds

$$R(\theta^\star) \leq \frac{m\cdot\hat{R}(\theta^\star)}{m-d_{cp}} + \left(\frac{\ln\binom{m}{d_{cp}}+\ln(m)+\ln\left(\frac{1}{\beta}\right)}{2(m-d_{cp})}\right)^{\frac{1}{2}} \tag{25}$$

For additional readings and recent advancements on compression-based bounds for both realizable and agnostic problems, the interested reader is refered to [97], [98], [99]. In [97], Hanneke and Kontorovich show that the optimal rates of agnostic compression schemes with compression rate k is often $\sqrt{k\ln(m/k)/m}$ which is in contrast with the known rate $\sqrt{d/m}$ of convergence of complexity-based agnostic problems, for a VC-dimension d. In [99], the authors studied stable compression schemes for a family of supervised learning algorithms. A new and enhanced margin bound for

Theorem	Learning theory	Data-indep. a-priori	Data-depen. a-posteriori	Data-realizable hard-constr.	Data-agnostic soft-constr.	Error bound dependence on
1	Scenario	✓	✗	✓	✗	$\epsilon(m, n_\theta, \beta)$
2	Scenario	✗	✓	✓	✗	$\epsilon(m, s^\star_m, \beta)$
3	Scenario	✗	✓	✓	✓	$\epsilon(m, s^\star_m, \beta)$
4	VC-Complexity	✓	✗	✓	✗	$\epsilon(m, d, \beta)$
5	Rademacher-Complex.	✗	✓	✗	✓	$\epsilon(m, r, \|w\|, \hat{R}, \beta)$
6	VC-Complexity	✗	✓	✗	✓	$\epsilon(m, d, \hat{R}, \beta)$
7	Compression	✓	✗	✓	✗	$\epsilon(m, d_{cp}, \beta)$
8	Compression	✗	✓	✗	✓	$\epsilon(m, d_{cp}, \hat{R}, \beta)$

TABLE I

A COMPARISON BETWEEN THE REVIEWED THEOREMS, FUNCTIONAL DEPENDENCY OF THE BOUNDS AND THEIR APPLICABILITY A-PRIORI, E.G, EVEN BEFORE SOLVING THE LEARNING PROBLEM, AND TO REALIZABLE LEARNING/AGNOSTIC PROBLEMS, I.E., OPTIMIZATION METHODS LEADING TO NULL/NON-NULL EMPIRICAL TRAINING ERROR.

SVM is proposed and removing a log factor. In [98], Cohen and Kontorovich discuss agnostic learning with unbounded metric losses and introduce a new technique called semi-stable compression.

V. Experimental Procedure and Results

A numerical procedure has been developed to compare the effectiveness of the revised bounds. The procedure works as follows:

1) **Initialize**: Choose the parameters for the bounds, including the kernel type, scale parameter, confidence level, number of experiments N_{exp} and ρ.
2) **Sample training data:** Randomly select one of the thirteen data sets, a subset of features (between 50% to 100% of n_x) and sample size (between 5% and 35% of the original size such that $m \leq 2000$ for efficiency's sake).
3) **Train an SVM model:** Solve program (6) and compute an unbiased estimator for the error probability $R(\theta^\star)$.
4) **Generalization bounds:** Compute the number of support vectors $s^\star_m$, VC dimension, the average of the ramp loss r, and $||w||$. Use $s^\star_m$, β and other statistics to compute the agnostic learning bounds as previously described. If any upper bound exceeds 1, set it to 1 (non-informative). If there are multiple bounds available, such as agnostic complexity-based ones, choose the bound with the smallest magnitude
5) **Assess tightness:** Evaluate the tightness of the scenario bounds in each experiment by estimating the difference $\gamma_{SB} = \epsilon(s^\star_m) - R(\theta^\star)$ and similarly compute the tightness of the complexity-based bounds (γ_{CX}) and the compression-based bounds (γ_{CP}).
6) **Statistical analysis of randomized experiments** Repeat steps 2 to 5 a total of N_{exp} times. Compare the expectation and variance of the tightness metrics to assess the relative performance of the different bounds.

Quantitative and qualitative analyzes of the tightness of the error bounds and convergence are analyzed for varying hyperparameters and on thirteen real-world data sets obtained from UCI, OpenML repositories, and MNIST.

A. The thirteen data sets

The 13 real-world data sets have been modified for binary classification and randomly sampled to generate new synthetic data sets. Here is a brief description of each classification problem:

1) **MNIST**: A data set of handwritten digits with 784 features per image (pixels). We classify 14000 samples of digits 1 and 3.
2) **LOL**: 9879 League of Legends game outcomes characterized by 38 input features to be classified as win or lose.
3) **Winequality**: It contains 4898 wine samples, $n_x = 11$ features, to be classified as good (score exceeding 5) or bad wines.
4) **Ionosphere**: A collection of $m = 351$ radar samples to be labelled as good or bad using $n_x = 38$ features.
5) **Abalone**: The goal of predicting the age of abalones from $n_x = 8$ biological characteristics. Abalones with over nine rings are labelled as old.
6) **Ailerons**: a binarized version with $n_x = 39$ features and $m = 13750$ samples.
7) **Spambase**: 4601 labelled e-mails with $n_x = 57$ features. The goal is to classify future SPAM emails.
8) **Data Eye**: The goal is to predict when a person's eyes are open/closed given 14780 continuous EEG measurements with 14 features (missing data and outliers have been removed).
9) **Postures**: 13600 samples with 8 coordinates from five hand postures ($X_{0:2}$, $Y_{0:2}$ and $Z_{0:1}$). Postures 1 and 3 are used for the binary classification task.
10) **Banknote authentication**: $m = 1372$ images with $n_x = 4$ feature taken from genuine and forged banknote specimens. The goal is to identify forged samples.
11) **Dota**: The goal is to classify the binary outcome of m =102944 game plays from 116 features.
12) **Monk Problem**: $m = 601$ samples and $n_X = 6$ features usable to predict the outcome of a logical formula.
13) **Gina Agnostic**: A data set for agnostic handwritten digit recognition that contains $m = 3468$ samples of two digits having $n_x = 970$ features (pixels).

B. Qualitative Comparison

This section analyzes the behaviour of the bounds when the hyperparameters are varied, particularly the scaling parameter of the kernel. One interesting observation is that when using a Gaussian kernel and in an agnostic setting, the VC dimension is infinite a priori. Hence, the margin bound in Eq. (21) must be used. Figure 2 illustrates the trend of the bounds on the League of Legends data set, and a similar trend has been observed in many other data sets as well. It is worth noting that the complexity-based bound, computed according to Theorem 5, failed to capture the trend of the test error estimate. On the other hand, the scenario bound followed the error much better. Figure 2 depicts this issue and shows a study of underlying statistics used to compute the bounds. The complexity-based bound incorporates the training loss $\sum_{i=1}^{m} l^{\star}_{SVM}(x_i, y_i)$ and the norm $|w^{\star}|$. The training loss initially increases as we over-fit the data, while the norm $|w^{\star}|$ closely tracks the test error. However, changes in $|w^{\star}|$ are much smaller than the changes in the training loss. Consequently, the margin bound underestimates the impact of having a smaller norm and fails to capture the trend of the test error. In other experiments, we have observed that both the scenario and complexity bounds follow the true error, however, we have never encountered a failure with the scenario bounds.

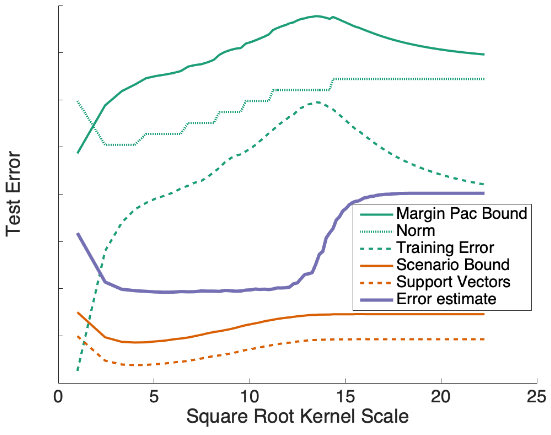

Fig. 2. The effect of the kernel scale on the test error, the formal generalization error bounds, and support vectors. The curves have been rescaled for clearer representation, as the focus is on observing the trends rather than the absolute values.

C. Quantitative Comparison

The experimental procedure described in Section V is applied to train N_{exp} =5000 linear soft-margin SVM classifiers for the 13 classification problems and setting kernels scale and ρ equal to one. This procedure yields 5000 randomized training sets containing $m \in [18, 2000]$ samples and $n_x \in [3, 970]$ features. The tightness scores of the formal generalization error bounds are compared for a confidence parameter $\beta = 10^{-2}$. The Rademacher margin bound in Eq. (21) is larger than one for most of the experiments (97.98%), whilst this happened in 2068 (41.36%) cases for the VC dimension-based bounds in Eq. (22). Hence, for linear SVMs, the VC bounds generally offer better results compared to the margin bounds. The compression-based bounds also resulted in a larger than one bound for many cases (87 % of the experiments). The scenario bounds resulted in better scores (tighter bounds) in about 71% of the 5000 experiments. Figure 3 compares the generalization error bounds with the test errors and presents the generalization error bounds and test errors from the 13 data sets with different coloured markers. As a figure of merit, a pair-wise comparison of the VC and scenario bounds is presented in the top left panel, whilst pair-wise comparisons of the empirical errors (x-axis) and the formal error bounds (y-axis) are displayed in the two panels on the right-hand side. Note that complexity-based bounds on the generalization error are always non-informative for the Gina agnostic, MNIST and Monk problem data sets, i.e., a generalization error in [0,1]. Differently, scenario bounds are often non-vacuous and always informative (and especially much better for the MINST data set).

1) Distribution of γ with respect to the individual data sets: We further investigate the efficacy of the bounds on the individual data sets and Figure 4 present the distribution of the tightness scores using Box and Whisker plots. Specifically, each box represents the 5th and 95th percentiles for the distribution of γ, calculated from randomized classification experiment (5000 for each data set) with linear soft-margin SVM and setting kernels scale and ρ equal to one. The top, middle, and bottom panels correspond to scenario bounds (γ_{SB}), complexity-based bounds (γ_{CX}), and compression bounds (γ_{CP}), respectively. The red regions indicate poor performance, with a discrepancy score $\gamma \geq 0.4$, and green regions represent good performing bounds with $\gamma < 0.4$. The analysis demonstrates that scenario bounds are often the most effective and useful in many classification problems, outperforming other formal generalization bounds for eight data sets. On the other hand, complexity-based bounds are effective only for three data sets (postures, data eye, and abalone), while compression bounds show good outcomes for two data sets (banknote and MNIST).

2) Bounds behaviour with respect to the test error: Figure 5 presents the formal error bounds and empirical error estimated from a test sets. The test errors are sorted in ascending order on the x-axis. The pentagram green markers represent the compression bounds, the blue markers denote the complexity-based bounds, and the red cross markers represent the scenario bounds. This analysis reveals that the scenario bounds have the best performance when the SVM model achieves high accuracy. However, when the accuracy of the SVM is low, e.g., an error greater than 0.2 in Fig 5 for the linear case, the statistical learning bounds can provide tighter generalization guarantees. It is also worth noting that for lower errors, the compression bounds occasionally outperform the complexity bounds but still do not perform as well as the scenario bounds.

3) Results for different ρ and optimized kernel scale: To analyze the effect of ρ and kernel type, we chose three values for ρ 1, 100, and 5000, and used two different kernel functions, linear and Gaussian. We trained 5000 classifiers for each of the six combinations of kernel type and regularization parameter. Unlike our previous analysis, we optimized the kernel scale using an empirical/heuristic optimization method (the MATLAB *auto* option). Note that as ρ approaches infinity, the soft-margin constraints in the SVM training programs

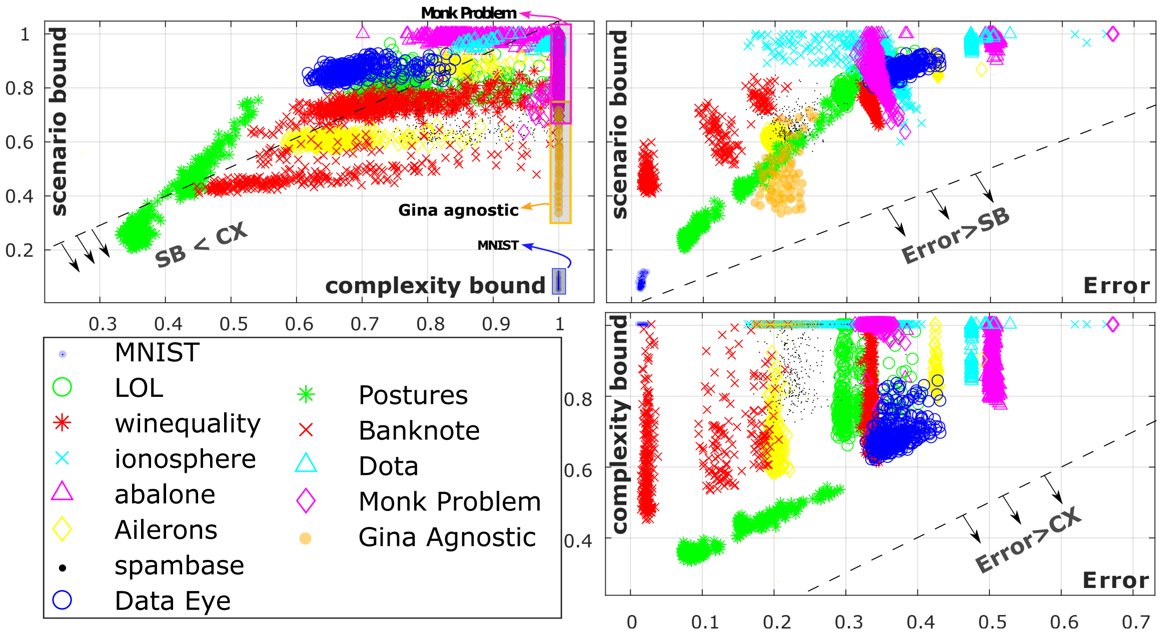

Fig. 3. Scenario and complexity-based bounds compared to the error estimate for soft-margin linear SVM classifiers. On the right panels, the dashed lines show the regions where the formal bound fails, whilst the left panels show experiments having scenario bounds tighter than the complexity bounds. The markers indicate different data sets.

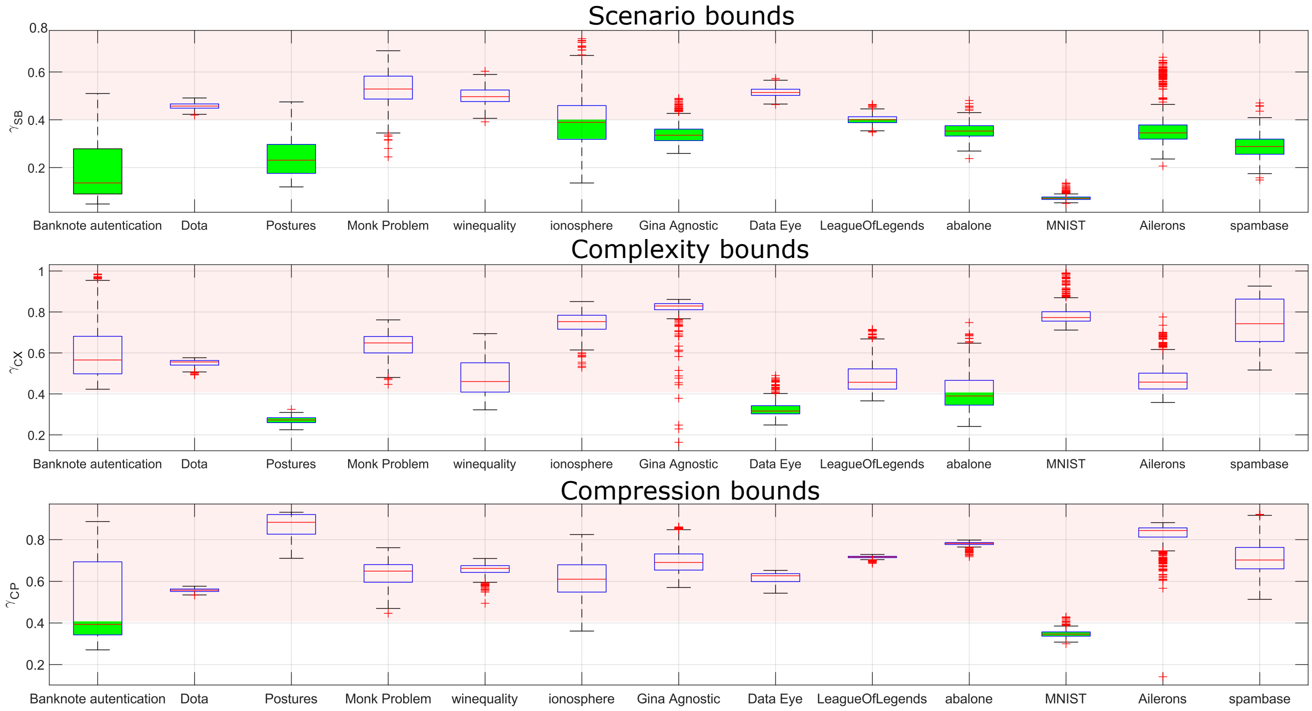

Fig. 4. Box and Whisker plots of tightness scores on the individual data sets. Scenario, complexity, and compression bound are presented, respectively, in the top, central and lower panels. In green colour, we highlight experiments with a relatively small discrepancy score, $\gamma < 0.4$, i.e., the tighter bounds. Clearly, scenario bounds outperformed the reviewed bounds in many cases.

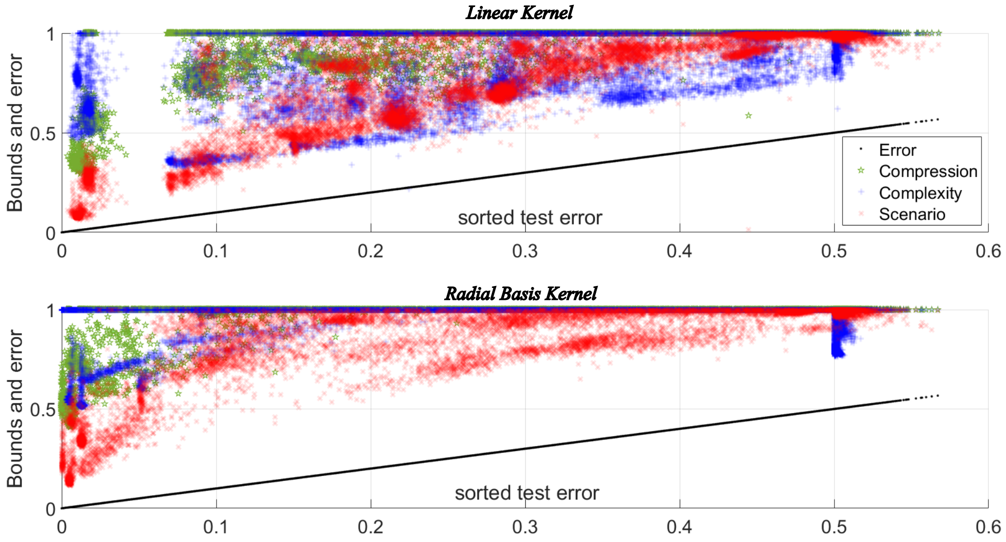

Fig. 5. The resulting scenario bounds (red markers), complexity bounds (blue markers) and compression-based bounds (green markers) are sorted by the values of the empirical test error estimates (black markers and x-axis). Results for the linear and Gaussian kernels are presented in the top and bottom panels respectively. Scenario bounds generally outperform other statistical learning bonds, especially when the error estimate is small.

	$\rho = 1$			$\rho = 100$			$\rho = 5000$		
Kernel	$\mathbb{E}[\gamma_{CX}] \pm \sigma$	$\mathbb{E}[\gamma_{SB}] \pm \sigma$	$\mathbb{E}[\gamma_{CP}] \pm \sigma$	$\mathbb{E}[\gamma_{CX}] \pm \sigma$	$\mathbb{E}[\gamma_{SB}] \pm \sigma$	$\mathbb{E}[\gamma_{CP}] \pm \sigma$	$\mathbb{E}[\gamma_{CX}] \pm \sigma$	$\mathbb{E}[\gamma_{SB}] \pm \sigma$	$\mathbb{E}[\gamma_{CP}] \pm \sigma$
Linear	0.46 ± 0.12	0.54± 0.10	0.61 ± 0.13	0.56 ± 0.18	0.45 ± 0.13	0.74 ± 0.13	0.57 ± 0.19	0.35 ± 0.14	0.64 ± 0.16
Gaussian	0.40 ± 0.10	0.49 ± 0.04	0.50± 0.06	0.67 ± 0.15	0.59 ± 0.13	0.70 ± 0.15	0.72 ± 0.13	0.49± 0.17	0.72 ± 0.16
	$[\underline{\gamma}_{CX}, \overline{\gamma}_{CX}]$	$[\underline{\gamma}_{SB}, \overline{\gamma}_{SB}]$	$[\underline{\gamma}_{CP}, \overline{\gamma}_{CP}]$	$[\underline{\gamma}_{CX}, \overline{\gamma}_{CX}]$	$[\underline{\gamma}_{SB}, \overline{\gamma}_{SB}]$	$[\underline{\gamma}_{CP}, \overline{\gamma}_{CP}]$	$[\underline{\gamma}_{CX}, \overline{\gamma}_{CX}]$	$[\underline{\gamma}_{SB}, \overline{\gamma}_{SB}]$	$[\underline{\gamma}_{CP}, \overline{\gamma}_{CP}]$
Linear	[0.11,0.87]	[0.29,0.86]	[0.43,0.91]	[0.15,0.98]	[0.15,0.80]	[0.42,0.98]	[0.05,0.74]	[0.07,0.99]	[0.14,0.93]
Gaussian	[0.25,0.99]	[0.34,0.99]	[0.43,0.99]	[0.48,0.99]	[0.28,0.92]	[0.45,0.98]	[0.46,1.0]	[0.10,0.92]	[0.40,0.99]

TABLE II

THE RESULTS FOR γ_{CX}, γ_{SB}, AND γ_{CP} FOR LINEAR AND GAUSSIAN KERNELS AND THREE VALUES OF THE REGULARIZATION PARAMETER ρ. RESULTS WERE OBTAINED FOR AN OPTIMIZED KERNEL SCALE (MATLAB 'AUTO' OPTION) AND 5000 RANDOM EXPERIMENTS.

revert back to the original hard-constrained formulation. This means that increasing ρ increases the cost of violation, which force the optimizer to reduce the average magnitude of margin violations. However, a reduction in the number of support vectors, it's not guaranteed. Table II summarizes the numerical results of our analysis. We present the averages of γ_{CX}, γ_{CP}, and γ_{SB}, denoted by $\mathbb{E}[\cdot]$, as well as their standard deviations σ. We also report the minimum and maximum values for each metric, denoted by $\underline{\gamma}_{SB}$, $\underline{\gamma}_{CX}$, $\underline{\gamma}_{CP}$, and $\overline{\gamma}_{SB}$, $\overline{\gamma}_{CX}$, $\overline{\gamma}_{CP}$, respectively, over the 5000 experiments. Our findings indicate that for $\rho \geq 100$, scenario bounds provide much superior results compared to other bounds. However, for lower ρ values, complexity-based methods achieve better bounds on average.

VI. Discussion and Conclusion

This work compared scenario theoretic, compression-learning, and complexity-based approaches to derive formal generalization error bounds for SVM classifiers. Despite this restriction on the model class, the revised theories apply to other models and, more generally, to data-driven agnostic and data-realizable learning algorithms. The most significant theorems and mathematical tools used for the proofs and derivations have been reviewed and discussed. Scenario-based bounds are often tailored to the specific learning problem and can be data-dependent or data-independent. The number of support scenarios, the support vectors for SVM models, affects the width of data-dependent scenario bounds, which are thus closely related to the concept of compression size in compression-learning theory. Differently, data-independent scenario bounds depend on the number of optimization parameters, which relates to the expressiveness of the learning model, and closely relates to model complexity and VC-dimension. We proposed a series of randomized training experiments to study and compare the tightness of scenario bounds with traditional statistical learning approaches, i.e., compression-learning and complexity-based bounds. We found that the scenario bounds are often tighter (especially for hard-margin cases and for low empirical errors) and that the margin bound can fail to capture the error trend for changing hyperparameters, we did not experience this with the bounds prescribed by other theories. Future research on theoretical generalization errors must focus on the following key issues:

- Investigate the relationship between data-independent and data-dependent error bounds for ad ML methods. For instance, recent works prove the potential of scenario bounds on game-theoretic models, Echo-State networks and other ML tools for data description, regression and prediction [100]–[102].
- There is a need to tackle a lack of theoretical understanding on how to provide tight and non-empirical error

bounds for agnostic and reliable problems for non-iid samples and using non-stationary probability spaces and data sequences, e.g., [103].
- Recent works showed that it is possible to study the generalization of deep learning models without the need for training nor testing data [20], hence suggesting the possibility of defining error guarantees from inherent structural properties of the trained model $f(\theta^\star)$. Future research efforts are needed to understand the interplay between the parameters and structure of a trained model $f(\theta^\star)$ and its capacity to compress the data and generalize to new examples.
- Finally, we wish to remark that, currently, it is a problem to derive non-vacuous bounds for complex systems such as neural networks [104], where it remains problematic to analyze them in classical frameworks [105]. Regarding this, it could be of particular interest to look at generalization bounds for non-convex agnostic learning problems [90].

Acknowledgment

The authors would like to acknowledge the following grant agreements The H2020-ECSEL project "101007319 AI-TWILIGHT", the project NWA.1160.18.238 ”PrimaVera”, and the ERC project ”N 758824 INFLUENCE".

The authors acknowledge IEEE Transactions on Neural Networks and Learning Systems for publishing the finalized version of this paper titled ”A Survey on Scenario Theory, Complexity, and Compression-Based Learning and Generalization" in Early Access on 13 September 2023, with the following details:

- DOI: 10.1109/TNNLS.2023.3308828
- Published in: IEEE Transactions on Neural Networks and Learning Systems (Early Access)
- Pages: 1 - 15
- Print ISSN: 2162-237X
- Electronic ISSN: 2162-2388
- PubMed ID: 37703153

References

[1] D. E. Knuth, "Big omicron and big omega and big theta," *ACM Sigact News*, vol. 8, no. 2, pp. 18–24, 1976.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, ser. Springer series in statistics. Springer, 2009.

[3] P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.

[4] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[5] A. Isaksson, M. Wallman, H. Göransson, and M. Gustafsson, "Cross-validation and bootstrapping are unreliable in small sample classification," *Pattern Recognition Letters*, vol. 29, no. 14, pp. 1960–1965, 2008.

[6] B. Efron and C. Stein, "The Jackknife Estimate of Variance," *The Annals of Statistics*, vol. 9, no. 3, pp. 586 – 596, 1981.

[7] Z. Shao and M. J. Er, "Efficient leave-one-out cross-validation-based regularized extreme learning machine," *Neurocomputing*, vol. 194, pp. 260–270, 2016.

[8] A. Onan, S. Korukoğlu, and H. Bulut, "Ensemble of keyword extraction methods and classifiers in text classification," *Expert Systems with Applications*, vol. 57, pp. 232–247, 2016.

[9] A. Onan, "Biomedical text categorization based on ensemble pruning and optimized topic modelling," *Computational and Mathematical Methods in Medicine*, vol. 2018, 2018.

[10] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Information Processing & Management*, vol. 53, no. 4, pp. 814–833, 2017.

[11] A. Onan and M. A. Toçoğlu, "A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification," *IEEE Access*, vol. 9, pp. 7701–7722, 2021.

[12] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[13] A. Dal Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit card fraud detection: A realistic modeling and a novel learning strategy," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 8, pp. 3784–3797, 2018.

[14] D. Zhang, J. Ma, J. Yi, X. Niu, and X. Xu, "An ensemble method for unbalanced sentiment classification," in *2015 11th International Conference on Natural Computation (ICNC)*, 2015, pp. 440–445.

[15] Z. Wang, M. Liu, Y. Cheng, and R. Wang, "Robustly fitting and forecasting dynamical data with electromagnetically coupled artificial neural network: A data compression method," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 6, pp. 1397–1410, 2017.

[16] J. Xu, W. Du, Y. Jin, W. He, and R. Cheng, "Ternary compression for communication-efficient federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 3, pp. 1162–1176, 2022.

[17] C. Zhang, D. Tao, T. Hu, and B. Liu, "Generalization bounds of multitask learning from perspective of vector-valued function learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 1906–1919, 2021.

[18] S. Liu, S. Xue, J. Wu, C. Zhou, J. Yang, Z. Li, and J. Cao, "Online active learning for drifting data streams," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 186–200, 2023.

[19] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban, "Metric learning from imbalanced data with generalization guarantees," *Pattern Recognition Letters*, vol. 133, pp. 298–304, 2020.

[20] C. Martin, T. Peng, and M. Mahoney, "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data," *Nature Communications*, vol. 12, p. 4122, 07 2021.

[21] X. Yue, M. Nouiehed, and R. Al Kontar, "SALR: Sharpness-aware learning rate scheduler for improved generalization," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[22] X. Wu, J. Zhang, and F.-Y. Wang, "Stability-based generalization analysis of distributed learning algorithms for big data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 3, pp. 801–812, 2020.

[23] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[24] O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy, "Proper learning, Helly number, and an optimal SVM bound," in *Annual Conference Computational Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 125, 2020, pp. 582–609.

[25] L. Oneto, N. Navarin, M. Donini, A. Sperduti, F. Aiolli, and D. Anguita, "Measuring the expressivity of graph kernels through statistical learning theory," *Neurocomputing*, vol. 268, pp. 4–16, 2017.

[26] L. Oneto, D. Anguita, and S. Ridella, "PAC-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis," *Pattern Recognition Letters*, vol. 80, pp. 200 – 207, 2016.

[27] S. Sun, J. Shawe-Taylor, and L. Mao, "PAC-bayes analysis of multi-view learning," *Information Fusion*, vol. 35, pp. 117–131, 2017.

[28] P. Germain, A. Habrard, F. Laviolette, and E. Morvant, "PAC-bayes and domain adaptation," *Neurocomputing*, vol. 379, pp. 379 – 397, 2020.

[29] W. Lee, H. Kim, and J. Lee, "Compact class-conditional domain invariant learning for multi-class domain adaptation," *Pattern Recognition*, vol. 112, p. 107763, 2021.

[30] L. Hellerstein and R. A. Servedio, "On PAC learning algorithms for rich boolean function classes," *Theoretical Computer Science*, vol. 384, no. 1, pp. 66–76, 2007.

[31] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[32] N. Littlestone and M. K. Warmuth, "Relating data compression and learnability," University of California Santa Cruz, Tech. Rep., 1986.

[33] M. C. Campi and S. Garatti, "Scenario optimization with relaxation: a new tool for design and application to machine learning problems," in *59th IEEE Conference on Decision and Control (CDC)*, 2020.

[34] R. Rocchetta, Q. Gao, and M. Petkovic, "Scenario-based generalization bound for anomaly detection support vector machine ensembles," in *Proceedings of the 30th ESREL Conference and 15th PSAM Conference*, 2020.

[35] R. Rocchetta, Q. Gao, D. Mavroeidis, and M. Petkovic, "A robust model selection framework for fault detection and system health monitoring with limited failure examples: Heterogeneous data fusion and formal sensitivity bounds," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 105140, 2022.

[36] R. Rocchetta, Q. Gao, and M. Petkovic, "Soft-constrained interval predictor models and epistemic reliability intervals: A new tool for uncertainty quantification with limited experimental data," *Mechanical Systems and Signal Processing*, vol. 161, p. 107973, 2021.

[37] F. Fabiani, K. Margellos, and P. J. Goulart, "On the robustness of equilibria in generalized aggregative games," in *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020, pp. 3725–3730.

[38] M. De Angelis, R. Rocchetta, A. Gray, and S. Ferson, "Constructing consonant predictive beliefs from data with scenario theory," in *International Symposium on Imprecise Probability: Theories and Applications*. PMLR, 2021, pp. 357–360.

[39] A. Carè, M. C. Campi, F. A. Ramponi, S. Garatti, and A. R. Cobbenhagen, "A study on majority-voting classifiers with guarantees on the probability of error," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1013–1018, 2020.

[40] R. Rocchetta, L. Crespo, and S. Kenny, "Solution of the benchmark control problem by scenario optimization," *Proceedings of the ASME Dynamic Systems and Control Conference, DSCC, October,*, 2019.

[41] R. Rocchetta and L. G. Crespo, "A scenario optimization approach to reliability-based and risk-based design: Soft-constrained modulation of failure probability bounds," *Reliability Engineering & System Safety*, vol. 216, p. 107900, 2021.

[42] A. Gray, A. Wimbush, M. de Angelis, P. Hristov, D. Calleja, E. Miralles-Dolz, and R. Rocchetta, "From inference to design: A comprehensive framework for uncertainty quantification in engineering with limited information," *Mechanical Systems and Signal Processing*, vol. 165, p. 108210, 2022.

[43] M. C. Campi and S. Garatti, "Compression, generalization and learning," in *arXiv*, 2023.

[44] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *J. ACM*, vol. 36, no. 4, p. 929–965, 1989.

[45] R. Yarullin and S. Obiedkov, "From equivalence queries to PAC learning: The case of implication theories," *International Journal of Approximate Reasoning*, vol. 127, pp. 1–16, 2020.

[46] O. David, S. Moran, and A. Yehudayoff, "Supervised learning through the lens of compression," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, 2016.

[47] H. Chase and J. Freitag, "Model theory and machine learning," *The Bulletin of Symbolic Logic*, vol. 25, no. 03, p. 319–332, 2019.

[48] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[49] K. Margellos, M. Prandini, and J. Lygeros, "On the connection between compression learning and scenario based single-stage and cascading optimization problems," *IEEE Transactions on Automatic Control*, vol. 60, no. 10, pp. 2716–2721, 2015.

[50] L. Romao, A. Papachristodoulou, and K. Margellos, "On the exact feasibility of convex scenario programs with discarded constraints," *IEEE Transactions on Automatic Control*, vol. 68, no. 4, pp. 1986–2001, 2023.

[51] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[52] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, pp. 199–222, 2004.

[53] H. Drucker, C. C, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Advances in Neural Information Processing Systems*, vol. 9, 11 2003.

[54] C. Suryanarayana, C. Sudheer, V. Mahammood, and B. Panigrahi, "An integrated wavelet-support vector machine for groundwater level prediction in Visakhapatnam, India," *Neurocomputing*, vol. 145, pp. 324–335, 2014.

[55] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 21, no. 6, pp. 2560 – 2574, 2007.

[56] S. Tian, J. Yu, and C. Yin, "Anomaly detection using support vector machines," in *Advances in Neural Networks – ISNN 2004*, 2004, pp. 592–597.

[57] A. Widodo and B.-S. Yang, "Machine health prognostics using survival probability and support vector machine," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8430 – 8437, 2011.

[58] D. Tax and R. Duin, "Support vector data description," *Machine Learning*, vol. 54, pp. 45–66, 01 2004.

[59] Y.-P. Zhao, G. Huang, Q.-K. Hu, and B. Li, "An improved weighted one-class support vector machine for turboshaft engine fault detection," *Engineering Applications of Artificial Intelligence*, vol. 94, p. 103796, 2020.

[60] R. Khemchandani, S. Chandra ***et al.***, "Twin support vector machines for pattern classification," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 29, no. 5, pp. 905–910, 2007.

[61] Q. Ye, C. Zhao, N. Ye, and X. Chen, "Localized twin SVM via convex minimization," *Neurocomputing*, vol. 74, no. 4, pp. 580–587, 2011.

[62] S. Sharma, R. Rastogi, and S. Chandra, "Large-scale twin parametric support vector machine using pinball loss function," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, pp. 987–1003, 2021.

[63] M. Tanveer, A. Sharma, and P. Suganthan, "Least squares KNN-based weighted multiclass twin SVM," *Neurocomputing*, 2020.

[64] X. Pan, Z. Yang, Y. Xu, and L. Wang, "Safe screening rules for accelerating twin support vector machine classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1876–1887, 2018.

[65] L. V. Utkin, "An imprecise extension of SVM-based machine learning models," *Neurocomputing*, vol. 331, pp. 18–32, 2019.

[66] M. Tanveer, T. Rajani, R. Rastogi, Y.-H. Shao, and M. Ganaie, "Comprehensive review on twin support vector machines," *Annals of Operations Research*, pp. 1–46, 2022.

[67] Z. Wang, Y.-H. Shao, L. Bai, and N.-Y. Deng, "Twin support vector machine for clustering," *IEEE Transactions on neural networks and learning systems*, vol. 26, no. 10, pp. 2583–2588, 2015.

[68] Z. Qi, Y. Tian, and Y. Shi, "Laplacian twin support vector machine for semi-supervised classification," *Neural Netw.*, vol. 35, no. C, p. 46–53, nov 2012.

[69] Y. Xu, Z. Yang, and X. Pan, "A novel twin support-vector machine with pinball loss," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 2, pp. 359–370, 2017.

[70] H. Yu, X. Yang, S. Zheng, and C. Sun, "Active learning from imbalanced data: A solution of online weighted extreme learning machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 4, pp. 1088–1103, 2019.

[71] J. Lou, Y. Jiang, Q. Shen, R. Wang, and Z. Li, "Probabilistic regularized extreme learning for robust modeling of traffic flow forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 4, pp. 1732–1741, 2023.

[72] P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien, "PAC-bayesian theory meets bayesian inference," 2017.

[73] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer, "PAC-bayesian inequalities for martingales," *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7086–7093, 2012.

[74] L. Ralaivola, M. Szafranski, and G. Stempfel, "Chromatic PAC-bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes," *The Journal of Machine Learning Research*, vol. 11, pp. 1927–1956, 2010.

[75] J. Chen, Z. Guo, H. Li, and C. L. P. Chen, "Regularizing scale-adaptive central moment sharpness for neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[76] V. N. Vapnik and A. Y. Chervonenkis, *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*. Springer International Publishing, 2015, pp. 11–30.

[77] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed., ser. Adaptive computation and machine learning. MIT Press, 2012.

[78] L. Oneto, A. Ghio, S. Ridella, and D. Anguita, "Local Rademacher complexity: Sharper risk bounds with and without unlabeled samples," *Neural Networks*, vol. 65, pp. 115–125, 2015.

[79] S. Floyd and M. Warmuth, "Sample compression, learnability, and the Vapnik-Chervonenkis dimension," *Machine learning*, vol. 21, no. 3, pp. 269–304, 1995.

[80] L. Yunwen, D. Ürün, Z. Ding-Xuan, and K. Marius, "Data-dependent generalization bounds for multi-class classification," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2995–3021, 2019.

[81] A. Cannon, J. M. Ettinger, D. Hush, and C. Scovel, "Machine learning with data dependent hypothesis classes," *J. Mach. Learn. Res.*, vol. 2, p. 335–358, mar 2002.

[82] A. Maalouf, G. Eini, B. Mussay, D. Feldman, and M. Osadchy, "A unified approach to coreset learning," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2022.

[83] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 254–263.

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

[85] G. Calafiore and M. Campi, "Uncertain convex programs: randomized solutions and confidence levels," *Mathematical Programming*, vol. 102, no. 1, pp. 25–46, Jan 2005.

[86] 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.

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

[88] A. Carè, S. Garatti, and M. C. Campi, "Scenario min-max optimization and the risk of empirical costs," *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2061–2080, 2015.

[89] P. Mohajerin 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, 2015.

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

[91] S. Garatti and M. C. Campi, "Risk and complexity in scenario optimization," *Mathematical Programming*, Nov 2019.

[92] R. Rocchetta, L. G. Crespo, and S. P. Kenny, "A scenario optimization approach to reliability-based design," *Reliability Engineering & System Safety*, vol. 196, p. 106755, 2020.

[93] S. Garatti and M. C. Campi, "The risk of making decisions from data through the lens of the scenario approach," *IFAC-PapersOnLine*, vol. 54, no. 7, pp. 607–612, 2021, 19th IFAC Symposium on System Identification SYSID 2021.

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

[95] S. Hanneke and A. Kontorovich, "Optimality of svm: Novel proofs and tighter bounds," *Theoretical Computer Science*, vol. 796, pp. 99 – 113, 2019.

[96] T. Graepel, R. Herbrich, and J. Shawe-Taylor, "PAC-bayesian compression bounds on the prediction error of learning algorithms for classification," *Machine Learning*, vol. 59, pp. 55–76, January 2005.

[97] S. Hanneke and A. Kontorovich, "A sharp lower bound for agnostic learning with sample compression schemes," in *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 98. PMLR, 22–24 Mar 2019, pp. 489–505.

[98] D. T. Cohen and A. Kontorovich, "Learning with metric losses," in *Proceedings of 35th Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 178, 02–05 Jul 2022, pp. 662–700.

[99] S. Hanneke and A. Kontorovich, "Stable sample compression schemes: New applications and an optimal svm margin bound," in *Algorithmic Learning Theory*. PMLR, 2021, pp. 697–721.

[100] F. Fele and K. Margellos, "Probably approximately correct Nash equilibrium learning," *IEEE Transactions on Automatic Control*, vol. 66, no. 9, pp. 4238–4245, 2021.

[101] L. B. Armenio, L. Fagiano, E. Terzi, M. Farina, and R. Scattolini, "Optimal training of echo state networks via scenario optimization," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 5183–5188, 2020.

[102] L. G. Crespo, B. K. Colbert, S. P. Kenny, and D. P. Giesy, "On the quantification of aleatory and epistemic uncertainty using sliced-normal distributions," *Systems & Control Letters*, vol. 134, p. 104560, 2019.

[103] S. Hanneke, "Learning whenever learning is possible: Universal learning under general stochastic processes," in *2020 Information Theory and Applications Workshop (ITA)*, 2020, pp. 1–95.

[104] R. Yang, R. Jia, X. Zhang, and M. Jin, "Certifiably robust neural ode with learning-based barrier function," *IEEE Control Systems Letters*, 2023.

[105] V. Nagarajan and J. Z. Kolter, "Uniform convergence may be unable to explain generalization in deep learning," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 11 615–11 626.

Roberto Rocchetta Dr. Roberto Rocchetta is a research associate at the University of Applied Sciences and Arts of Southern Switzerland (SUPSI), affiliated with the DACD-ISAAC intelligent energy system group and a member of the NCCR-automation group. Dr. Rocchetta obtained his PhD in Energy and Reliability Engineering in 2019 and has since worked at the National Institute of Aerospace, USA, at the Technical University of Eindhoven Netherlands. His research focuses on decision-making under risk and uncertainty, artificial intelligence and machine learning for energy system optimization and reliability engineering.

Alexandr Mey Dr. Alexander Mey is a Postdoctoral researcher at the Department of Mathematics and Computer Science at TU/Eindhoven. He received his PhD with a focus on data science and machine learning in (2019), and his research interests focus on reinforcement learning, decision-making and probabilistic causality.

Frans A. Oliehoek Dr. Frans A. Oliehoek is Associate Professor at Delft University of Technology, where he leads a group on interactive learning and decision making, is one of the scientific directors of the Mercury machine learning lab, and is director and co-founder of the ELLIS Unit Delft. He received his Ph.D. in Computer Science (2010) from the University of Amsterdam (UvA), and held positions at various universities including MIT, Maastricht University and the University of Liverpool. Frans' research interests revolve around intelligent systems that learn about their environment via interaction, building on techniques from machine learning, AI and game theory. He has served as PC/SPC/AC at top-tier venues in AI and machine learning, and currently serves as associate editor for JAIR and AIJ. He is a Senior Member of AAAI, and was awarded a number of personal research grants, including a prestigious ERC Starting Grant.