Journal of Machine Learning Research 22 (2021) 1-38 Submitted 6/21; Revised 10/21; Published 11/21

A Theory of the Risk for Optimization with Relaxation and its Application to Support Vector Machines

Marco C. Campi MARCO.CAMPI@UNIBS.IT
Department of Information Engineering
University of Brescia
via Branze 38, 25123 Brescia, Italy

Simone Garatti SIMONE.GARATTI@POLIMI.IT
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
piazza L. da Vinci 32, 20133 Milano, Italy

Editor: Corinna Cortes

Abstract

In this paper we consider optimization with relaxation, an ample paradigm to make data-driven designs. This approach was previously considered by the same authors of this work in Garatti and Campi (2019), a study that revealed a deep-seated connection between two concepts: *risk* (probability of not satisfying a new, out-of-sample, constraint) and *complexity* (according to a definition introduced in paper Garatti and Campi, 2019). This connection was shown to have profound implications in applications because it implied that the risk can be estimated from the complexity, a quantity that can be measured from the data without any knowledge of the data-generation mechanism. In the present work we establish new results. First, we expand the scope of Garatti and Campi (2019) so as to embrace a more general setup that covers various algorithms in machine learning. Then, we study classical support vector methods – including SVM (Support Vector Machine), SVR (Support Vector Regression) and SVDD (Support Vector Data Description) – and derive new results for the ability of these methods to generalize. All results are valid for any finite size of the data set. When the sample size tends to infinity, we establish the unprecedented result that the risk approaches the ratio between the complexity and the cardinality of the data sample, regardless of the value of the complexity.

Keywords: optimization, optimization with relaxation, generalization, risk quantification, support vector machines

1. Introduction

Various techniques in machine learning – and more generally in data-driven decision-making – hinge upon the following two ingredients:

(i) a cost function $c(x)$, which one would like to make as small as possible;
(ii) constraints $f(x, \delta_i) \leq 0$, where δ_i are observations.

In the process of optimizing the cost $c(x)$, constraints $f(x, \delta_i) \leq 0$ can be accounted for in various ways. A flexible paradigm – which contains more rigid setups as extreme cases – is obtained by relaxing the constraints and make them "soft" according to the following

scheme

$$\begin{aligned} \min_{\substack{x\in\mathcal{X}\\ \xi_i\geq 0, i=1,\ldots,N}} \quad & c(x)+\rho\sum_{i=1}^{N}\xi_i \\ \text{subject to:} \quad & f(x,\delta_i)\leq \xi_i, \quad i=1\ldots,N. \end{aligned} \tag{1}$$

The interpretation of (1) is that some constraints $f(x,\delta_i)\leq 0$ can be violated for the purpose of improving the cost value, but constraints violation has itself a cost as expressed by the auxiliary optimization variables ξ_i: if $\xi_i>0$, then constraint $f(x,\delta_i)\leq 0$ is relaxed to $f(x,\delta_i)\leq\xi_i$ and this generates the regret ξ_i, which adds to the original cost $c(x)$. The parameter ρ is used to set a suitable trade-off between the original cost and the cost generated by the regret for violating constraints.

In machine learning, optimization with constraints relaxation plays a major role in various contexts and we provide below examples taken from the gallery of support vector methods, which is a main focus of attention in the present paper.

Example 1 (Support Vector Regression - SVR) *Let $\{\delta_i\}_{i=1}^N=\{(\mathbf{u}_i,y_i)\}_{i=1}^N$ be a training set, where the $\mathbf{u}_i$'s are instances living in a suitable input domain, for example $\mathbb{R}^n$, and the y_i's are the corresponding output values in $\mathbb{R}$. For given parameters $\tau,\rho>0$, one considers the optimization program (see e.g. Schölkopf et al., 1998):*

$$\begin{aligned} \min_{\substack{w,\gamma\geq 0, b\in\mathbb{R}\\ \xi_i\geq 0, i=1,\ldots,N}} \quad & (\gamma+\tau\|w\|^2)+\rho\sum_{i=1}^{N}\xi_i \\ \text{subject to:} \quad & |y_i-\langle w,\mathbf{u}_i\rangle-b|-\gamma\leq\xi_i, \quad i=1,\ldots,N. \end{aligned} \tag{2}$$

The cost function in (2) minimizes a weighted sum of the size γ of the "tube" used for prediction and the regularization term $\|w\|^2$, to which penalties ξ_i are added for output measurements y_i that are not in the tube (i.e., their distance from the interpolating function $\langle w,\mathbf{u}_i\rangle+b$ is more than γ). Upon solving program (2), one finds the solution $(w^*,\gamma^*,b^*,\xi_i^*)$, which gives the prediction tube

$$|y-\langle w^*,\mathbf{u}\rangle-b^*|\leq\gamma^*. \tag{3}$$

When a new value $\bar{\mathbf{u}}$ is received, the corresponding output $\bar{y}$ is forecast to be in the tube, that is, in the range of values of y that satisfy the relation $|y-\langle w^*,\bar{\mathbf{u}}\rangle-b^*|\leq\gamma^*$ and one incurs a prediction error if $\bar{y}$ happens not to belong to this range. ⋆

Example 2 (Support Vector Data Description - SVDD) *This is an example of an un-supervised learning technique. Let $\{\delta_i\}_{i=1}^N=\{\mathbf{p}_i\}_{i=1}^N$ be a set of points in $\mathbb{R}^n$. SVDD constructs a sphere whose center c^* and radius γ^* are obtained from program (see e.g. Tax and Duin, 2004):*

$$\begin{aligned} \min_{\substack{c,\gamma\geq 0\\ \xi_i\geq 0, i=1,\ldots,N}} \quad & \gamma+\rho\sum_{i=1}^{N}\xi_i \\ \text{subject to:} \quad & \|\mathbf{p}_i-c\|^2-\gamma\leq\xi_i, \quad i=1,\ldots,N. \end{aligned} \tag{4}$$

One next out-of-sample point is predicted to be in the sphere and an error is incurred if this does not happen. ⋆

SVR and SVDD can be cast more generally than in the above examples by referring to kernel approaches able to lift the working domain into a feature space. SVR, SVDD, as well as Support Vector Machines (SVM) are studied in detail in Section 3, where we apply our new risk theory to derive tight evaluations for the probability of error of these machines. More generally, problem (1) accommodates methods that arise in numerous contexts in data science where relaxation of the constraints can be used to tone down the importance of anomalous observations (sometimes called *outliers*) that would otherwise generate ill-designed solutions, while in other cases relaxation is even strictly necessary to circumvent infeasibility issues (like in SVM with non-linearly separable data). Our theory here developed applies to all these cases.

As previously mentioned, program (1) furnishes a flexible scheme that allows the designer to explore various prospective solutions obtained as ρ varies between the two extremes $\rho = 0$ (no regret for constraints violation) and $\rho = \infty$ (infinite regret for constraints violation, in which case all constraints are rigidly enforced). In this process of selection, one is aided by quantitative tools that describe the quality of the solutions x^*. Recalling (i) and (ii), it is natural that the designer is concerned about the achieved cost $c(x^*)$ and the ensuing *risk* $V(x^*)$, where, for a generic value of the optimization variable x, the risk

$$V(x) = \mathbb{P}\{\delta : f(x, \delta) > 0\}$$

($\mathbb{P}$ is the probability that governs the generation of δ values) quantifies the probabilistic level of constraints violation (in SVR, violating a constraint corresponds to providing an interval for $\bar{y}$ that does not include its actual value, while in SVDD it amounts to construct a sphere that does not contain the next point). One key-aspect worth noticing is that $c(x^*)$ becomes readily available to the designer after the solution x^* to problem (1) has been computed; in contrast, the risk of x^* cannot be directly evaluated since its definition involves $\mathbb{P}$, which is normally not available to the user. Hence, evaluating $V(x^*)$ requires to develop solid theoretical results and this is instrumental to boost a general trust in data-driven methods, especially in contexts where data are used in automated designs, and not just as a simple support to decisions. The ultimate goal of this contribution is to put forward a new theory that holds true *distribution-free* and yet it allows for tight and practically useful evaluations of the risk.

1.1 Previous results this paper builds upon

In Garatti and Campi (2019), the problem of estimating $V(x^*)$ was addressed in a convex setup ($c(x)$ and $f(x, \delta)$ are convex in x) by adopting the so-called wait-&-judge perspective of Campi and Garatti (2018). Specifically, a certificate on $V(x^*)$ is obtained from the value taken by an observable quantity s^*, called *complexity* and defined as the number of δ_i's for which $f(x^*, \delta_i) \geq 0$ (i.e., s^* = no. of active constraints + no. of violated constraints). Interestingly, the solution x^* can be fully reconstructed from the constraints appearing in the definition of s^* and, therefore, s^* can be interpreted as the complexity of representation

of the solution. More discussion on this point is provided in Section 2.1, where we also offer a systematic and detailed comparison of the results of the present paper with other approaches in the literature.

As is intuitive, the number of violated constraints alone (which, when divided by the number of scenarios, gives the *empirical risk*) is not a valid indicator of the true risk $V(x^*)$ since optimization generates a bias towards larger risks by drifting the solution against the constraints. The main achievement of Garatti and Campi (2019) consists in showing that the complexity is instead strictly linked to $V(x^*)$ and, as such, it can be used to accurately judge the level of risk. This discovery implies a profound and revealing truth: two solutions with the same empirical risk can have quite different true risks $V(x^*)$ depending on hidden mechanisms sitting in the method; nonetheless, it is a universal fact that all these mechanisms are captured by the complexity, which, alone, offers an accessible door to evaluate the risk.

Very importantly, applying this theory requires no model for how observations δ_i are generated. As a matter of fact, although δ_i are modeled as random outcomes from a probability distribution $\mathbb{P}$, the obtained results apply irrespective of $\mathbb{P}$, and $\mathbb{P}$ remains undefined throughout the algorithmic and theoretical developments of the method. This is practically important since in many applications assuming that $\mathbb{P}$ is known to the designer is unrealistic: $\mathbb{P}$ refers to the "real world" and can be a truly complex object in modern data science for which hardly complete a-priori knowledge is available (think e.g. of biological or social systems, or of problems arising in autonomous driving, just to cite but a few examples).

To better frame the above mentioned result, we also indicate that the work Garatti and Campi (2019) follows in the wake of the so-called "scenario approach", initiated with the seminal paper Calafiore and Campi (2005) and then continued in a stream of theoretical developments, Campi and Garatti (2008); Schildbach et al. (2013); Margellos et al. (2014); Zhang et al. (2015); Carè et al. (2015), with application to fields like control system design, Calafiore and Campi (2006); Schildbach et al. (2014); Grammatico et al. (2016); Falsone et al. (2019), system identification, Welsh and Rojas (2009); Campi et al. (2009); Welsh and Kong (2011); Crespo et al. (2014, 2015, 2016); Lacerda and Crespo (2017); Garatti et al. (2019), and learning, Campi (2010); Campi and Carè (2013); Margellos et al. (2015); Caré et al. (2018).

1.2 New contributions of this paper

Building upon the achievements of Garatti and Campi (2019), in this paper we establish new results.

(a) We consider the important class of support vector methods, which have been developed in machine learning for classification and regression problems. In support vector regression methods, the dichotomy between cost and constraints satisfaction described above corresponds to the dichotomy between having informative regressors or classifiers and their probability of misprediction. One contribution of this paper is to establish all the connections between the general theory of Garatti and Campi (2019) and support vector methods, including the nontrivial adaptation of the theory to the specific setups when required. It is then shown how the new theory allows for a more

reliable usage of support vector methods, especially in relation to the long-standing problem of tuning hyper-parameters, which is key to obtain good solutions.

(b) Support vector methods are studied in Section 3. For a better understanding of this part, we will first revisit in Section 2 the theory of Garatti and Campi (2019) and we will present it in a broader setup than that of Garatti and Campi (2019) by considering convex optimization over generic (possibly infinite dimensional) vector spaces. This is a necessary step since generic vector spaces is the natural setup for support vector methods whenever the so called kernel trick is applied. Exploiting the full power of the theory of Garatti and Campi (2019) in a general setup is a second contribution of the present paper.[1]

(c) We provide asymptotic characterizations of the risk evaluations in (b) and, as a corollary of our theory, Section 2.1 establishes for the first time that the risk of the solution tends in great generality to the ratio between the complexity of the solution and the sample size N, as $N \to \infty$. While our main thrust in this paper remains that of establishing tight evaluations of the risk that are usable for any finite sample size N, we remark that this convergence result is unprecedented and sheds new light on the existence of empirical observables that allow one to obtain estimates that converge to the true risk. This new achievement outdoes known results based on measures of complexity of the class of hypotheses as well as results obtained in the domain of compression schemes.

2. Risk Assessment in Optimization with Constraints Relaxation

In this section, we revisit and extend the theory of Garatti and Campi (2019) for the assessment of $V(x^*)$ (Theorem 1). We start by formally stating three assumptions. The first specifies the mathematical frame of work, and the second requires that x^* is well-defined. The third assumption is instead a technical requirement whose implications will be commented upon later.

Assumption 1 (mathematical setup) *x is an element of a vector space $\mathcal{X}$ (possibly infinite dimensional). $c(x)$ and, for any given δ, $f(x,\delta)$ are convex functionals of x. The scenarios δ_i, $i = 1,\dots,N$, form an independent and identically distributed (i.i.d.) random sample from a probability space $(\Delta, \mathcal{D}, \mathbb{P})$, that is, $\delta_1,\dots,\delta_N$ is an outcome from the probability space $(\Delta^N, \mathcal{D}^N, \mathbb{P}^N)$, where $\mathcal{D}^N = \mathcal{D}\otimes\cdots\otimes\mathcal{D}$ and $\mathbb{P}^N = \mathbb{P}\times\cdots\times\mathbb{P}$ are the product σ-algebra and the product probability measure, respectively.* $\star$

Assumption 2 (existence and uniqueness) *Consider optimization problems as in* (1) *where N is substituted with any index $m = 0, 1, \dots$ (i.e., m is any nonnegative integer) and δ_i, $i = 1,\dots,m$, is an i.i.d. sample from $(\Delta, \mathcal{D}, \mathbb{P})$. For every m and for every outcome $(\delta_1, \delta_2, \dots, \delta_m)$, it is assumed that these optimization problems admit a solution (i.e., the problems are feasible and the infimum is achieved on the feasibility set). If for one of*

1. Note that working in infinite dimensional spaces rules out the possibility of using results where the complexity is *a-priori* bounded by the dimension of the optimization vector as is done, e.g., in Campi and Garatti (2008).

these optimization problems more than one solution exists, one solution is singled out by the application of a convex tie-break rule, which breaks the tie by minimizing an additional convex functional $t_1(x)$, and, possibly, other convex functionals $t_2(x)$, $t_3(x)$, ... if the tie still occurs.[2] ⋆

The following is a technical non-accumulation assumption of functionals $f(x, \delta)$.

Assumption 3 (non-accumulation) *For every x in $\mathcal{X}$, $\mathbb{P}\{\delta : f(x, \delta) = 0\} = 0$.* ⋆

This assumption is linked to the concept of non-degeneracy introduced in Definition 3 of Garatti and Campi (2019) and it is often satisfied when δ itself does not accumulate (e.g., when it has a density). Moreover, this assumption is a reasonable modeling simplification even when δ is discrete but a fine-grained quantity. Examples and more discussion will be provided in Section 3 in connection to support vector methods.

The following theorem is a reformulation in the present context of the main result of Garatti and Campi (2019).

Theorem 1 *For a given value in $(0, 1)$ of the confidence parameter β, consider for any $k = 0, 1, \ldots, N-1$ the polynomial equation in the t variable*

$$\binom{N}{k} t^{N-k} - \frac{\beta}{2N} \sum_{i=k}^{N-1} \binom{i}{k} t^{i-k} - \frac{\beta}{6N} \sum_{i=N+1}^{4N} \binom{i}{k} t^{i-k} = 0, \tag{5}$$

and, for $k = N$, consider the polynomial equation in the t variable

$$1 - \frac{\beta}{6N} \sum_{i=N+1}^{4N} \binom{i}{N} t^{i-N} = 0. \tag{6}$$

For any $k = 0, 1, \ldots, N-1$, equation (5) has exactly two solutions in $[0, +\infty)$, which we denote with $\underline{t}(k)$ and $\overline{t}(k)$ ($\underline{t}(k) \leq \overline{t}(k)$). Instead, equation (6) has only one solution in $[0, +\infty)$, which we denote with $\overline{t}(N)$, while we define $\underline{t}(N) = 0$. Let $\underline{\epsilon}(k) := \max\{0, 1 - \overline{t}(k)\}$ and $\overline{\epsilon}(k) := 1 - \underline{t}(k)$, $k = 0, 1, \ldots, N$. Under Assumptions 1, 2 and 3, for any Δ and $\mathbb{P}$ it holds that

$$\mathbb{P}^N\{\underline{\epsilon}(s^*) \leq V(x^*) \leq \overline{\epsilon}(s^*)\} \geq 1 - \beta, \tag{7}$$

where x^* is the solution to (1), possibly after breaking the tie according to Assumption 2, and s^* is the number of δ_i's for which $f(x^*, \delta_i) \geq 0$. ⋆

Proof The proof is easily obtained by noticing that the proof of Theorem 4 in Garatti and Campi (2019), given for the case of optimization over Euclidean spaces, applies *mutatis mutandis* to the present more general setup. Details are simple and left to the reader. ■

The main message conveyed by Theorem 1 is that it is possible to construct an interval $[\underline{\epsilon}(s^*), \overline{\epsilon}(s^*)]$ where $V(x^*)$ lies with high confidence $1 - \beta$, and no information on Δ and

2. Note that only the tie with respect to x is broken by $t_1(x)$, $t_2(x)$, $t_3(x)$, ... On the other hand, for a given x^* the values of ξ_i, $i = 1, \ldots, m$, remain unambiguously determined at optimum by relation $\xi_i^* = f(x^*, \delta_i)$, so that no tie on ξ_i, $i = 1, \ldots, m$, can persist after the tie on x is broken.

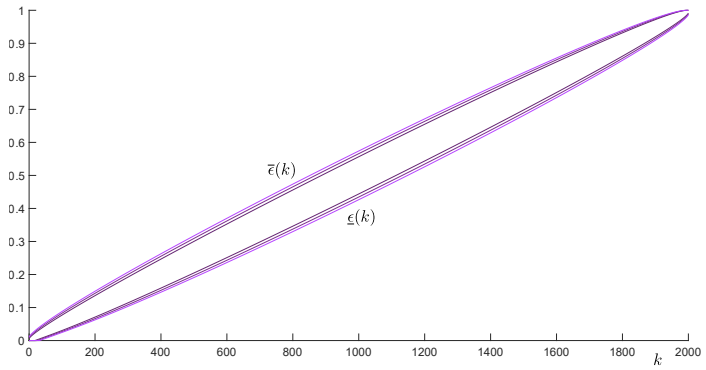

Figure 1: $\underline{\epsilon}(k)$ and $\overline{\epsilon}(k)$ for $N = 2000$ and $\beta = 10^{-4}, 10^{-6}, 10^{-8}$. As β decreases, the intervals gently enlarge.

$\mathbb{P}$ is required in this process of evaluation (distribution-free result). The interval depends on s^*, which is an observable that can be computed from the data record $\delta_1, \ldots, \delta_N$, and for different values of s^* one obtains different ranges for $V(x^*)$, showing that s^* carries fundamental information for the estimation of $V(x^*)$. Figure 1 depicts $\underline{\epsilon}(k)$ and $\overline{\epsilon}(k)$ for $N = 2000$ and $\beta = 10^{-4}, 10^{-6}, 10^{-8}$, from which we see that small and informative intervals are obtained even for extremely high levels of confidence. Further building on the result in Theorem 1, in Section 2.1 we shall provide asymptotic evaluations that establish the fact that the risk tends to the ratio between the complexity and the sample size N as N tends to infinity.[3]

The typical usage of Theorem 1 is as follows. The designer solves (1) repeatedly for various values of ρ and obtains various solutions x^* achieving different trade-offs between cost and risk. As ρ varies, the cost is computed, while Theorem 1 allows one to bound the risk based on the observed value of the complexity s^*. In this way, the designer can generate a cost-risk plot like the one depicted in Figure 2, where the cost $c(x^*)$ and the interval $[\underline{\epsilon}(s^*), \overline{\epsilon}(s^*)]$ for $V(x^*)$ are depicted corresponding to various values of s^* (this plot refers to a numerical example presented in Section 4). The user is thus provided with the relevant information to select the solution that achieves the best compromise for the problem at hand. This same reasoning can be carried over to other hyper-parameters besides ρ appearing in the optimization program. As an example, in Section 4 we shall consider the tuning of the hyper-parameters of a Gaussian kernel.

Remark 1 *A common practice to estimate the risk of the solution consists in dividing the available observations into a training sample δ_i, $i = 1, \ldots, N_T$, which is used to compute a solution x_T^*, and a validation sample δ_i, $i = N_T+1, \ldots, N_T+N_V$, with $N_T+N_V = N$, by which $V(x_T^*)$ is estimated from the ratio*

$$\frac{\textit{no. of } \delta_i, i = N_T+1, \ldots, N_T+N_V, \textit{ such that } f(x_T^*, \delta_i) > 0}{N_V}. \tag{8}$$

3. One reason for the tightness of the results in this paper is that these results focus on the risk of the solution rather than being uniform with respect to all potential solutions. This sets an important departure from uniform theories based on the Vapnik-Chervonenkis dimension.

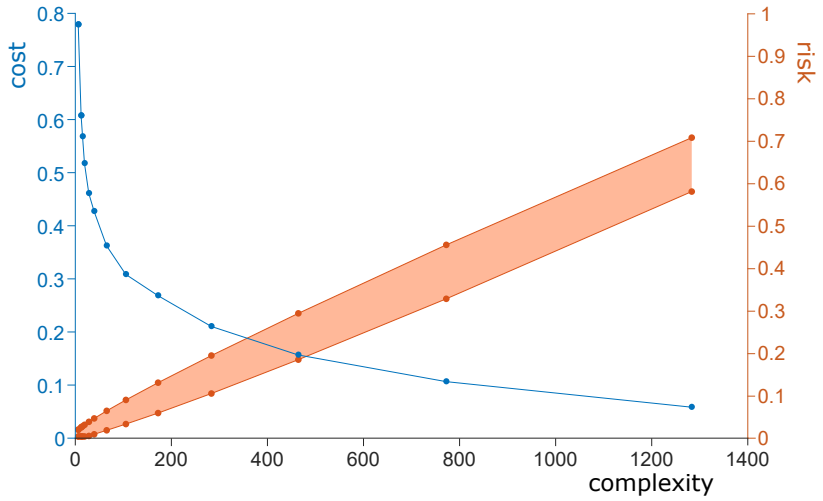

Figure 2: The cost-risk plot. Dots in the picture correspond to the values k of s^* that have been observed for a range of selections of the parameter ρ. The decreasing function indicates the cost while the intervals show the range for the risk.

This way of proceeding is justified by the law of large numbers, which ensures that the above ratio tends to $V(x^*_T)$ asymptotically, and evaluations of the estimation accuracy can be formulated for any finite N_V as well. However, validation requires sacrificing a portion of the observations to estimate the risk, rather than using them for design purposes. This can be difficult to accept in applications where the data are a valuable and scarce resource. Beyond this point, we feel advisable to offer two more comments that clarify some important theoretical aspects. (i) Using all data to design only moderately reduces the power of the data to achieve the dual effect of generating useful estimates of the risk (refer to our finite sample results and the asymptotic theory in Section 2.1); (ii) One may be tempted to infer $V(x^*)$ (the risk of the solution obtained from all scenarios) from an equation like (8) *applied to a solution obtained from a subset of the data (sometimes, it is also suggested to repeatedly apply* (8) *over, say,* 10% *leave-out schemes and average the results). However, this way of proceeding is not only invalid from a statistical point of view, it can also generate highly imprecise evaluations when the solution is subject to stochastic variability after leaving out some of the data points.*[4] *Instead, Theorem 1 provides a well-principled and statistically valid framework to estimate the risk $V(x^*)$, with no waste of information for the design of the solution.* $\star$

Remark 2 *For the sake of precision, we feel advisable to point out some extra details in relation to the selection of the hyper-parameter ρ. Each single use of Theorem 1 has a probability β of providing an incorrect evaluation of the risk. Hence, when the theorem is repeatedly applied to obtain evaluations corresponding to various values of ρ, say p values,*

4. This is also true in cross-validation schemes for hyper-parameter selection and is one of the reasons why, besides validation data, one is recommended to save further test data for a final evaluation of the chosen solution.

the probability that the evaluation is wrong in at least one of the values is upper bounded by $p \cdot \beta$*. Since we cannot exclude that the user hits a wrong evaluation whenever one exists, the overall procedure is guaranteed with probability* $1 - p \cdot \beta$*. In other words, using the cost-risk plot to select the value of* ρ *may result in that the risk is not in the computed interval in at most one case out of* $1/(p \cdot \beta)$*. On the other hand, as pointed out in Section 2.1, enforcing very small values of* β *is "cheap" (i.e., it requires moderate numbers of data points) and therefore compensating for the extra probability owing to the increase from* β *to* $p \cdot \beta$ *is of little concern for practical purposes.* ⋆

Remark 3 *It is worth noticing that the result in Theorem 1 has some connection with the theory of conformal prediction of Vovk et al. (2005) and Shafer and Vovk (2008) (see also Lei et al., 2013, Vovk, 2013, and Györfi and Walk, 2019 for other contributions). Specifically, it can be shown that the notion of constraint violation for the solution* x^* *implicitly introduces a conformity measure over the scenarios and the risk* $V(x^*)$ *can then be interpreted as a (training sample) conditional coverage. However, rather than using the general tools of conformal prediction that work best for the* mean *of the conditional coverage (also known as unconditional coverage, see Vovk, 2013), Theorem 1 leverages the specific structure of* (1) *to provide tight characterizations of the* distribution *of the conditional coverage. This allows one to obtain estimators of* $V(x^*)$ *without resorting to calibration procedures, as done in Vovk (2013).* ⋆

Remark 4 (The price of knowledge) *Interestingly, reversing the order of the narration, one obtains a new, stimulating, interpretation of Theorem 1. Theorem 1 claims that small complexity (compared to the sample size* N*) implies reliable models and, viceversa, reliable models requires small complexity. As previously mentioned, the complexity is the size of the sub-sample of observations from which one can re-construct the model. In other words, all other observations become unimportant, and can be discarded, for modeling purposes once this sub-sample is known. A model is our means to describe reality, and it embodies our knowledge about how the portion of reality we are interested in acts and reacts to external stimuli. Hence, Theorem 1 can be interpreted that reliable knowledge within the scheme of* (1) *can only exist in the presence of an abundance of ineffectual observations. This is what we call the* price of knowledge.

This reasoning applies broadly to learning schemes that somehow relate to the refutation theory of Popper's philosophy of science, Popper (1962). Suppose that we construct a model led by a principle of parsimony (for example we build the smallest regression layer limited by straight lines in $\mathbb{R}^2$ *that contains two-dimensional points whose coordinates are height and weight of observed members of a population). If a new observation agrees with the model (e.g., (height,weight) of a new member falls in the layer), we take it as a confirmation of the model, otherwise, if the new observation does not agree with the model, the model is invalidated (or "refuted") and a new model able to accommodate the new observation, besides all observations previously collected, is put in its place. Along this process, if* s^* *remains a small fraction of the total number of observations, the model becomes "corroborated" and is expected to survive new invalidation tests as they come along down the stream of observations. In the context we are describing here, the model agrees with all observations (in our formalization, this corresponds to take* $\rho = \infty$*) and, if we assume that the non-accumulation*

assumption holds (in e.g. the case of the layer, this follows from requiring that the distribution of points has a density – see Section 3.1 – an assumption that approximately applies in the case of large populations), then we can use Theorem 1 to draw precise and quantitative results supporting the expectation that the model becomes "corroborated". Exploring the connections between the mathematical results provided in this paper and broad themes about inductive methods as expressed in the philosophical literature is a goal of great breadth that certainly deserves a much closer attention than that given to it here. ⋆

2.1 Discussion, asymptotic results, and comparison with the existing literature

Our result, Theorem 1, can be cast within the frame of work of compression schemes, Floyd and Warmuth (1995). In this context, a natural baseline of comparison is Graepel et al. (2005) in which the so-called luckiness function was introduced to adjust the risk to the value of an observable similarly to our complexity paradigm. Specifically, the result of Graepel et al. (2005) (somehow adapted to the framework of our paper) can be summarized as follows. Denote by c^* the number of the smallest set of δ_i's that is sufficient to reconstruct x^*. As is clear, these δ_i's are a subset of the δ_i's for which $f(x^*, \delta_i) \geq 0$ and, thanks to Assumption 3, they include with probability one all the δ_i's for which $f(x^*, \delta_i) = 0$. Moreover, let r^* be the number of δ_i's for which $f(x^*, \delta_i) > 0$. Note that $s^* \leq c^* + r^* \leq 2s^*$ with probability one (it can be that $c^* + r^*$ is strictly greater than s^* because a δ_i can be counted twice as an observation which is needed to reconstruct x^* as well as an observation for which $f(x^*, \delta_i) > 0$). Then, Theorems 2 and 3 in Graepel et al. (2005) show that the function

$$\tilde{\epsilon}(c, r) = \begin{cases} \frac{\ln \binom{N}{c} + \ln N + \ln \frac{1}{\beta}}{N-c}, & c = 0, 1, \ldots, N, \quad r = 0, \\ \frac{r}{N-c} + \sqrt{\frac{\ln \binom{N}{c} + 2\ln N + \ln \frac{1}{\beta}}{2(N-c)}}, & c = 0, 1, \ldots, N, \quad r = 1, \ldots, N \end{cases} \tag{9}$$

evaluated in (c^*, r^*) can be used to upper bound $V(x^*)$ according to formula

$$\mathbb{P}^N\{V(x^*) \leq \tilde{\epsilon}(c^*, r^*)\} \geq 1 - \beta. \tag{10}$$

Below, we compare $\tilde{\epsilon}(c^*, r^*)$ with our $\overline{\epsilon}(s^*)$ appearing in equation (7). Before doing so, however, we feel advisable to note that having also a lower bound as in our (7) is a guarantee of tightness of our result beyond its comparison with the result provided by (9) and (10). Tables 1 – 4 show the value of $\tilde{\epsilon}(c, r)$ against that of $\underline{\epsilon}(c + r)$ and $\overline{\epsilon}(c + r)$ for various values of c and r.[5]

We next move to study the asymptotic behavior of our bounds $\overline{\epsilon}(k)$ and $\underline{\epsilon}(k)$ as $N \to \infty$.

Theorem 2 *Functions $\underline{\epsilon}(k)$ and $\overline{\epsilon}(k)$ introduced in Theorem 1 are subject to the following bounds:*

$$\overline{\epsilon}(k) \leq \frac{k}{N} + C\frac{\sqrt{k}\ln\frac{1}{\beta} + \sqrt{k}\ln k + 1}{N} \tag{11}$$

$$\underline{\epsilon}(k) \geq \frac{k}{N} - C\frac{\sqrt{k}\ln\frac{1}{\beta} + \sqrt{k}\ln k + 1}{N} \tag{12}$$

5. Note that this comparison is sharp for situations for which $s^* = c^* + r^*$. It is fair noticing that in many cases $s^* < c^* + r^*$, which introduces a further source of conservatism in the usage of (10).

$r\backslash c$	10	20	40	80	160	320
0	0 - 0.036	0 - 0.052	0.013 - 0.08	0.04 - 0.13	0.11 - 0.23	0.25 - 0.4
50	0.028 - 0.11	0.036 - 0.12	0.051 - 0.15	0.082 - 0.19	0.15 - 0.28	0.29 - 0.45
100	0.066 - 0.17	0.074 - 0.18	0.09 - 0.2	0.12 - 0.25	0.19 - 0.34	0.34 - 0.51
150	0.11 - 0.23	0.11 - 0.24	0.13 - 0.26	0.17 - 0.31	0.24 - 0.39	0.39 - 0.56
200	0.15 - 0.28	0.16 - 0.3	0.17 - 0.32	0.21 - 0.36	0.28 - 0.44	0.43 - 0.61
250	0.19 - 0.34	0.20 - 0.35	0.22 - 0.37	0.25 - 0.41	0.33 - 0.49	0.48 - 0.65

Table 1: Values of $\underline{\epsilon}(c+r)$ and $\overline{\epsilon}(c+r)$; $N = 1000$, $\beta = 10^{-5}$.

$r\backslash c$	10	20	40	80	160	320
0	0.073	0.116	0.191	0.319	0.54	0.944
50	0.251	0.299	0.367	0.459	0.584	0.764
100	0.301	0.351	0.419	0.513	0.643	0.837
150	0.352	0.402	0.471	0.568	0.703	0.911
200	0.402	0.453	0.523	0.622	0.762	0.985
250	0.453	0.504	0.575	0.676	0.822	1

Table 2: Values of $\tilde{\epsilon}(c, r)$; $N = 1000$, $\beta = 10^{-5}$.

$r\backslash c$	10	20	40	80	160	320
0	0 - 0.018	0 - 0.026	0.007 - 0.041	0.021 - 0.067	0.052 - 0.12	0.12 - 0.21
50	0.014 - 0.055	0.018 - 0.061	0.025 - 0.074	0.04 - 0.098	0.073 - 0.15	0.14 - 0.23
100	0.033 - 0.086	0.036 - 0.092	0.044 - 0.10	0.06 - 0.13	0.09 - 0.17	0.16 - 0.26
150	0.052 - 0.12	0.056 - 0.12	0.065 - 0.13	0.081 - 0.16	0.12 - 0.2	0.19 - 0.29
200	0.073 - 0.15	0.077 - 0.15	0.085 - 0.16	0.1 - 0.19	0.14 - 0.23	0.21 - 0.32
250	0.094 - 0.17	0.098 - 0.18	0.11 - 0.19	0.12 - 0.21	0.16 - 0.26	0.23 - 0.34

Table 3: Values of $\underline{\epsilon}(c+r)$ and $\overline{\epsilon}(c+r)$; $N = 2000$, $\beta = 10^{-5}$.

$r\backslash c$	10	20	40	80	160	320
0	0.04	0.065	0.108	0.183	0.31	0.533
50	0.174	0.211	0.262	0.332	0.425	0.548
100	0.199	0.236	0.288	0.358	0.452	0.578
150	0.224	0.261	0.314	0.384	0.479	0.607
200	0.249	0.287	0.339	0.41	0.506	0.637
250	0.274	0.312	0.365	0.436	0.533	0.667

Table 4: Values of $\tilde{\epsilon}(c, r)$; $N = 2000$, $\beta = 10^{-5}$.

where C is a suitable constant (independent of k, N and β) and the bounds hold for $1 \leq k \leq N$ and $\beta \in (0,1)$, while, for $k = 0$, we have $\overline{\epsilon}(0) \leq (\ln(1/\beta) + 1) \cdot C/N$ and $\underline{\epsilon}(0) \geq 0$. $\star$

Proof See Appendix A. ■

In (11) and (12), the dependence in β is inversely logarithmic, which shows that "confidence is cheap". For any fixed k, we see that $\overline{\epsilon}(k)$ and $\underline{\epsilon}(k)$ merge onto the same value k/N as fast as $O(1/N)$, while for k that grows at the same rate as N, say $k = \mu N$, convergence towards k/N takes place at a rate $O(\ln(N)/\sqrt{N})$. Hence, we see that we can construct a strip around k/N whose size goes to zero as $O(\ln(N)/\sqrt{N})$ and the bi-variate distribution of risk and complexity all lies in the strip but a slim tail that expands beyond the strip whose probability is no more than β.

Going back to the comparison between (9), (10) and Theorem 1, this time in relation to asymptotic results, suppose first that $r^* = 0$ with probability one (no constraints violation - realizable case), in which case it holds that $s^* = c^*$ with probability one. This allows for a simple comparison between $\tilde{\epsilon}(c,0)$ and the upper bound for $\overline{\epsilon}(k)$ in equation (11), from which the following conclusions can be drawn:

(i) for any given β and for fixed value of c and k, we have that both $\tilde{\epsilon}(c,0)$ and $\overline{\epsilon}(k)$ converge to 0 as $N \to \infty$. Hence, both results capture the fact that the risk of x^* goes to 0 asymptotically when the complexity s^* keeps bounded as the number N of observations increases. The rate of convergence is slightly different and $\tilde{\epsilon}(c,0)$ converges to 0 as $O(\ln(N)/N)$ while $\overline{\epsilon}(k)$ converges to zero as $O(1/N)$. Provably, $O(1/N)$ is the fastest possible rate of convergence, Hanneke and Kontorovich (2019). It is worth mentioning that a bound converging to 0 as $O(1/N)$ was also obtained in Bousquet et al. (2020). This bound, however, differently from $\tilde{\epsilon}(r,0)$ and $\overline{\epsilon}(k)$ is only valid under the assumption that s^* is uniformly upper bounded for any N;

(ii) when c and k grow at the same rate as N, say $c = k = \mu N$, the behavior of $\overline{\epsilon}(k)$ and $\tilde{\epsilon}(c,0)$ are quite different. $\overline{\epsilon}(k)$ (and also $\underline{\epsilon}(k)$) converge to $k/N = \mu$ as $N \to \infty$ at a rate $O(\ln(N)/\sqrt{N})$. Instead, $\tilde{\epsilon}(c,0)$ does not converge to $c/N = \mu$ as $N \to \infty$, since, as shown in Appendix B, asymptotically it holds that $\tilde{\epsilon}(c,0) \geq 1 - (1 - c/N)(c/N)^{\frac{c/N}{1-c/N}}$ (which is bigger than $c/N = \mu$). This substantially different behavior can be also appreciated in Figure 3, where $\tilde{\epsilon}(c,0)$ along with $1-(1-c/N)(c/N)^{\frac{c/N}{1-c/N}}$ and $\overline{\epsilon}(k)$ and $\underline{\epsilon}(k)$ along with k/N are plotted as functions of $c = 0, 1, \ldots, N$ and of $k = 0, 1 \ldots, N$ for increasing values of N. Here, we see that the actual profile of $\tilde{\epsilon}(c,0)$ departs significantly above $1-(1-c/N)(c/N)^{\frac{c/N}{1-c/N}}$ for large values of c (in fact the only reason for pointing out the lower limit $1-(1-c/N)(c/N)^{\frac{c/N}{1-c/N}}$ was to show that convergence to c/N was missing). The result that $1 - \beta$ of the probabilistic mass of the bi-variate distribution of the risk and the complexity lies in the "lenticular" strip shown in Figure 3 whose width shrinks to zero at a rate $O(\ln(N)/\sqrt{N})$ is unprecedented and shows for the first time that in all situations (including those where the complexity is not

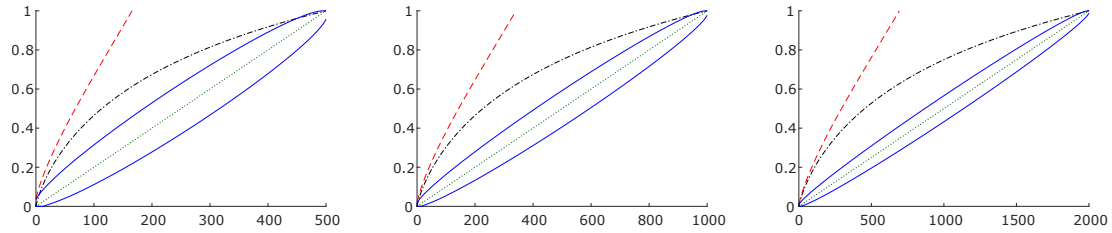

Figure 3: $\tilde{\epsilon}(c,0)$ (red dashed line), $1-(1-c/N)(c/N)^{\frac{c/N}{1-c/N}}$ (black dash-dotted line), $\overline{\epsilon}(k)$ and $\underline{\epsilon}(k)$ (blue solid lines), and k/N (green dotted line) as functions of $c = 0, 1, \ldots, N$ and of $k = 0, 1 \ldots, N$ for $N = 500$, $N = 1000$, and $N = 2000$.

a-priori upper bounded, a condition that applies e.g. to Support Vector methods with kernels) the ratio s^*/N is a consistent estimator of the risk $V(x^*)$.

For the case $r^* \neq 0$, similar to the discussion above one can conclude that $\tilde{\epsilon}(c,r)$ does not converge to $(c+r)/N$ in situations where c grows unbounded. In contrast, in the theory of the present paper the fact that $r^* \neq 0$ is not explicitly considered as a separate case and the complexity s^* accommodates all situations. Within this theory, the fact that for $k = \mu N$ both $\overline{\epsilon}(k)$ and $\underline{\epsilon}(k)$ converge to $k/N = \mu$ for all μ at a rate $O(\ln(N)/\sqrt{N})$ shows that s^* is an observable by which $V(x^*)$ can be consistently estimated in all situations and it reveals that the distinction between c^* and r^* is fictitious for the purpose of evaluating the risk.

3. Application to Support Vector Methods

In this section, the general theory for optimization with constraints relaxation is applied to various well known support vector methods. The results stemming from this analysis are unprecedented and show that complexity carries fundamental information to tightly judge the ability of these machines to generalize.

We consider in turn: SVR (Support Vector Regression), SVDD (Support Vector Data Description) and SVM (Support Vector Machine). To SVR and SVDD the theoretical apparatus developed in the previous section can be directly applied, while SVM requires some additional effort to rigorously accommodate some degenerate situations; the analysis for SVM also shows the versatility of the theory.

3.1 Support Vector Regression - SVR

Let $\{\delta_i\}_{i=1}^N = \{(\mathbf{u}_i, y_i)\}_{i=1}^N$ be a data set, where the $\mathbf{u}_i$'s are elements of a Hilbert space $\mathcal{U}$ and the y_i's are the corresponding output values in $\mathbb{R}$. Each data point is drawn independently of the others from a common probability distribution $\mathbb{P}$.

Remark 5 *Depending on the application, values $\mathbf{u}_i$ can be thought of as raw measurements of physical quantities or rather as measurements lifted into a feature space by means of a feature map $\varphi(\cdot)$, so that $\mathbf{u}_i = \varphi(\mathbf{m}_i)$, where $\mathbf{m}_i$ is a vector of measured quantities. Interestingly, when SVR is applied, the actual computation of the solution involves only the evaluation of inner products in the feature space, that is, $\langle \varphi(\mathbf{m}_k), \varphi(\mathbf{m}_j) \rangle$, which can be*

done without explicitly evaluating $\varphi(\mathbf{m}_i)$. Indeed, one can define a "kernel" $k(\mathbf{m}_k, \mathbf{m}_j) := \langle \varphi(\mathbf{m}_k), \varphi(\mathbf{m}_j) \rangle$ and working with function $k(\cdot, \cdot)$ enables one to implicitly operate in the (high-dimensional) feature space without ever computing explicitly the coordinates of the measurements in the lifted feature space. This is the so-called "kernel trick". Pushing all this even further, it can be observed that for the operation of the method one does not even need to provide an explicit description of the inner product $\langle \cdot, \cdot \rangle$ and of the feature map $\varphi(\cdot)$ from which $k(\cdot, \cdot)$ is defined by composition: in fact one can start off by assigning $k(\cdot, \cdot)$ directly and theoretical results in RKHS – Reproducing Kernel Hilbert Spaces – assure that this always corresponds to allocate a suitable couple $\langle \cdot, \cdot \rangle$ and $\varphi(\cdot)$ so that $k(\cdot, \cdot) = \langle \varphi(\cdot), \varphi(\cdot) \rangle$, provided that the kernel is positive definite (i.e., $\sum_{i=1}^{n} \sum_{j=1}^{n} k(\mathbf{m}_i, \mathbf{m}_j) c_i c_j \geq 0$, for all choices of n and all finite sequences of points $(\mathbf{m}_1, \ldots, \mathbf{m}_n)$ and real values $(c_1, \ldots, c_n)$). When adopting this standpoint, the interpretation of $k(\cdot, \cdot)$ is that it is a user-specified similarity function over pairs of data points in raw representation. $\star$

In the following, we refer to SVR with adjustable size as described in Schölkopf et al. (1998). For given parameters $\tau, \rho > 0$, consider the following optimization program, which, for easy reference, we repeat from the introduction (we here also specify more precisely the domain of optimization)[6]

$$\begin{aligned} \min_{\substack{w \in \mathcal{U}, \gamma \geq 0, b \in \mathbb{R} \\ \xi_i \geq 0, i=1,\ldots,N}} \quad & (\gamma + \tau \|w\|^2) + \rho \sum_{i=1}^{N} \xi_i \\ \text{subject to:} \quad & |y_i - \langle w, \mathbf{u}_i \rangle - b| - \gamma \leq \xi_i, \quad i = 1, \ldots, N. \end{aligned} \tag{13}$$

In this context the risk is interpreted as the probability of an erroneous prediction (i.e., a new observation $(\bar{u}, \bar{y})$ is not in the tube (3) constructed by SVR). Also, notice that the size of the prediction tube is known from the solution of the optimization program, while the theory here developed provides a fundamental grasp on the other relevant quantity, the probability of an erroneous prediction. These two pieces of information form the beacon to select a suitable value of the tuning parameter ρ. See also Section 4 for a numerical example.

Remark 6 *It is perhaps worth re-writing program* (13) *in terms of the original measurements when one resorts to a kernel lifting. As we show below, w^* is always given by a linear combination of the points $\mathbf{u}_i$, so that in* (13) *one can only consider solutions of the type $w = \sum_j \alpha_j \mathbf{u}_j$. Thus, one obtains $\|w\|^2 = \langle \sum_j \alpha_j \mathbf{u}_j, \sum_k \alpha_k \mathbf{u}_k \rangle = \sum_{j,k} \alpha_j \alpha_k \langle \mathbf{u}_j, \mathbf{u}_k \rangle$, while the constraints can be re-written as $|y_i - \sum_j \alpha_j \langle \mathbf{u}_j, \mathbf{u}_i \rangle - b| - \gamma \leq \xi_i$, $i = 1, \ldots, N$. Thus, in kernel notation, program* (13) *becomes*

$$\begin{aligned} \min_{\substack{\alpha_j, j=1,\ldots,N, \gamma \geq 0, b \in \mathbb{R} \\ \xi_i \geq 0, i=1,\ldots,N}} \quad & \gamma + \tau \sum_{j,k} \alpha_j \alpha_k k(\mathbf{m}_j, \mathbf{m}_k) + \rho \sum_{i=1}^{N} \xi_i \\ \text{subject to:} \quad & |y_i - \sum_j \alpha_j k(\mathbf{m}_j, \mathbf{m}_i) - b| - \gamma \leq \xi_i, \quad i = 1, \ldots, N, \end{aligned} \tag{14}$$

6. The above SVR formulation is suitable when the data are homoskedastic. Other convex formulations exist to model heteroskedastic processes, see Crespo et al. (2016).

which is a quadratic program. Often, (14) *is solved by resorting to its dual formulation. Letting* α_j^*, $j = 1, \ldots, N$, γ^*, b^*, *and* ξ_i^*, $i = 1, \ldots, N$ *be the optimal solution to* (14), *the tube in* (3) *becomes*

$$|y - \sum_j \alpha_j^* k(\mathbf{m}_j, \mathbf{m}) - b^*| \leq \gamma^*.$$

$\star$

Throughout, we make the following assumption.

Assumption 4 *Over the support of* $\mathbf{u}$, *the conditional distribution of* y *given* $\mathbf{u}$ *admits density.* $\star$

In order to apply the theory from Section 2 we need to show that the solution to (13) exists and is unique (Assumption 2) and that a non-accumulation assumption applies (Assumption 3). The validity of these facts is shown in the following.

Existence: While w belongs to a possibly infinite dimensional Hilbert space $\mathcal{U}$, the minimization problem in (13) (with m in place of N as required in Assumption 2) can be seen as finite dimensional because allowing for components of w outside the finite dimensional span of points $\mathbf{u}_i$, $i = 1, \ldots, m$, does not help satisfy the constraints (note that in the constraints w shows up only under the sign of inner product $\langle w, \mathbf{u}_i \rangle$), while it increases the cost function (write $w = w_{\mathbf{u}} + w_{\mathbf{u}}^{\perp}$, with $w_{\mathbf{u}} \in$ span of $\mathbf{u}_i$, $i = 1, \ldots, m$, and $w_{\mathbf{u}}^{\perp}$ orthogonal to the same span, and then apply Pitagora's theorem: $\|w\|^2 = \|w_{\mathbf{u}}\|^2 + \|w_{\mathbf{u}}^{\perp}\|^2$). Hence, (13) is a finite-dimensional problem with closed constraints and quadratic non-negative cost over the optimization domain. As such, it certainly admits solution.

Uniqueness: At optimum, w^* is certainly unique because, assuming by contradiction that there are two optimal solutions $(w_1^*, \gamma_1^*, b_1^*, \xi_{i,1}^*)$ and $(w_2^*, \gamma_2^*, b_2^*, \xi_{i,2}^*)$ with $w_1^* \neq w_2^*$, then an easy computation shows that the point half way between these two solutions would be feasible and superoptimal (the reader may also want to refer to Theorem 3 in Burges and Crisp (1999) where the same issue is discussed in relation to an algorithmically slightly different, but conceptually identical, problem). Instead, γ^*, b^* and ξ_i^* might be non-unique. To identify a unique solution we select the smallest γ^* and the b^* with smallest absolute value. Note that this certainly breaks the tie because the smallest γ^* is obviously unique while, if one had two values for b^* smallest in absolute value, say $b^* = \pm\bar{b}$, corresponding to the solutions $(w^*, \gamma^*, \bar{b}, \xi_{i,1}^*)$ and $(w^*, \gamma^*, -\bar{b}, \xi_{i,2}^*)$ (recall that w^* and γ^* must be the same at optimum), then optimality of these two solutions would imply that $\sum_{i=1}^N \xi_{i,1}^* = \sum_{i=1}^N \xi_{i,2}^*$ and therefore the solution half way between $(w^*, \gamma^*, \bar{b}, \xi_{i,1})$ and $(w^*, \gamma^*, -\bar{b}, \xi_{i,2})$, i.e., $(w^*, \gamma^*, 0, 0.5 \cdot \xi_{i,1} + 0.5 \cdot \xi_{i,2})$, would be feasible thanks to convexity, it would achieve the same cost as the other two solutions, but it would be preferred because it carries a smaller value for $|b^*|$ than in the two alleged solutions. Once w^*, γ^* and b^* are uniquely determined, also the ξ_i^*'s remain determined, see the footnote at the end of Assumption 2.

Non-accumulation: Non-accumulation requires that, $\forall w, \gamma, b$, one has:

$$\mathbb{P}\{|y - \langle w, \mathbf{u}\rangle - b| - \gamma = 0\} = 0.$$

Since the conditional distribution of y given $\mathbf{u}$ admits density, one has $\mathbb{P}\{|y-\langle w, \mathbf{u}\rangle-b|-\gamma = 0\} = \mathbb{P}\{\mathbb{P}\{|y - \langle w, \mathbf{u}\rangle - b| - \gamma = 0|\mathbf{u}\}\} = \mathbb{P}\{\mathbb{P}\{y = \langle w, \mathbf{u}\rangle + b \pm \gamma|\mathbf{u}\}\} = 0$.

Since all conditions are satisfied, we can apply Theorem 1 to SVR, which gives the following result.

Theorem 3 (Reliability of SVR) *With $\underline{\epsilon}(\cdot)$ and $\overline{\epsilon}(\cdot)$ as defined in Theorem 1, we have*

$$\mathbb{P}^N\{\underline{\epsilon}(s^*) \leq \mathbb{P}\{(\mathbf{u}, y) : |y - \langle w^*, \mathbf{u}\rangle - b^*| > \gamma^*\} \leq \overline{\epsilon}(s^*)\} \geq 1 - \beta,$$

where s^* is the number of $(\mathbf{u}_i, y_i)$'s for which $|y_i - \langle w^*, \mathbf{u_i}\rangle - b^*| \geq \gamma^*$. ⋆

3.2 Support Vector Data Description - SVDD

Support Vector Data Description is a data-driven technique used to identify a portion of space that covers most of the probabilistic mass from which data have been generated, while including little superfluous space. SVDD creates a spherically shaped form and, analogous to SVR, it can be made more flexible by lifting the data into a feature space so as to obtain more complex geometries in the original measurement space. See e.g. Tax and Duin (2004) for a more comprehensive description. See also Crespo et al. (2019) for an approach that allows one to apply SVDD to more complex geometries when working directly in the measurement space.

Let $\{\delta_i\}_{i=1}^N = \{\mathbf{p}_i\}_{i=1}^N$ be an independent data set in a Hilbert space $\mathcal{P}$ drawn from a common probability distribution $\mathbb{P}$. These points can be raw data or, in complete analogy with the discussion in Remark 5, data lifted into a feature space by means of a map $\varphi(\cdot)$. SVDD constructs a sphere in $\mathcal{P}$ by solving the following optimization program:

$$\begin{array}{ll} \min\limits_{\substack{c\in\mathcal{P},\gamma\geq 0\\ \xi_i\geq 0, i=1,\ldots,N}} & \gamma + \rho\displaystyle\sum_{i=1}^N \xi_i \\ \text{subject to:} & \|\mathbf{p}_i - c\|^2 - \gamma \leq \xi_i, \quad i = 1, \ldots, N. \end{array} \tag{15}$$

Remark 7 *When the original data are lifted into a feature space ($\mathbf{p} = \varphi(\mathbf{m})$) defined through a kernel, considering that the solution takes the form $c = \sum_j \alpha_j \mathbf{p_j}$ (see below), program* (15) *can be re-written as*

$$\begin{array}{ll} \min\limits_{\substack{\alpha_j, j=1,\ldots,N, \gamma\geq 0\\ \xi_i\geq 0, i=1,\ldots,N}} & \gamma + \rho\displaystyle\sum_{i=1}^N \xi_i \\ \text{subject to:} & k(\mathbf{m}_i, \mathbf{m}_i) + \displaystyle\sum_{j,k} \alpha_j\alpha_k k(\mathbf{m}_j, \mathbf{m}_k) - 2\sum_j \alpha_j k(\mathbf{m}_i, \mathbf{m}_j) - \gamma \leq \xi_i, \quad i = 1, \ldots, N. \end{array} \tag{16}$$

Moreover, the region in the measurement space obtained by the optimization procedure is given by

$$\Big\{\mathbf{m}: \quad k(\mathbf{m},\mathbf{m}) + \sum_{j,k} \alpha_j^* \alpha_k^* k(\mathbf{m}_j,\mathbf{m}_k) - 2\sum_j \alpha_j^* k(\mathbf{m},\mathbf{m}_j) \leq \gamma^*\Big\}.$$

⋆

We next address existence, uniqueness and non-accumulation for this problem.

Existence: Similarly to SVR, the optimal c^* must belong to the finite dimensional space generated by $\mathbf{p}_i$, $i = 1,\ldots,m$, and a solution to (15) certainly exists.

Uniqueness: At optimum, the center of the sphere c^* is unique while γ^* and the ξ_i^*'s may not be unique, refer to Theorems 2 and 3 in Wang et al. (2011); moreover non-uniqueness may only occur when $\rho = 1/M$ for some integer M, refer again to Theorem 3 in Wang et al. (2011). To break the tie if it occurs, select the smallest γ^*; note that in this way also the ξ_i^*'s remain uniquely determined as explained in the footnote at the end of Assumption 2.

Non-accumulation: For SVDD, non-accumulation requires the following condition to hold.

Assumption 5 *For any c and γ it holds that*

$$\mathbb{P}\{\|\mathbf{p} - c\|^2 = \gamma\} = 0. \tag{17}$$

⋆

This condition simply requires that probabilistic mass does not accumulate over hyper-spheres.

We now have the following theorem.

Theorem 4 (Reliability of SVDD) *With $\underline{\epsilon}(\cdot)$ and $\overline{\epsilon}(\cdot)$ as defined in Theorem 1, we have*

$$\mathbb{P}^N\{\underline{\epsilon}(s^*) \leq \mathbb{P}\{\mathbf{p}: \|\mathbf{p} - c^*\|^2 > \gamma^*\} \leq \overline{\epsilon}(s^*)\} \geq 1 - \beta,$$

where s^* is the number of $\mathbf{p}_i$'s for which $\|\mathbf{p}_i - c^*\|^2 \geq \gamma^*$. ⋆

3.3 Support Vector Machines - SVM

SVM is a well-known technique that constructs binary classifiers from a data set. Given a new out-of-sample case, the classifier predicts its label to be -1 or 1. -1 and 1 represent two different classes, whose meaning depends on the application at hand and can e.g. be *sick* or *healthy*, *right* or *wrong*, *male* or *female*. Among the vast literature on SVM, refer e.g. to Cortes and Vapnik (1995); Schölkopf and Smola (1998).

Let $\{\delta_i\}_{i=1}^N = \{(\mathbf{u}_i, y_i)\}_{i=1}^N$ be a data set of independent observations from a common probability distribution $\mathbb{P}$, where the $\mathbf{u}_i$'s are elements of a Hilbert space $\mathcal{U}$ and the y_i's

are the corresponding labels, -1 or 1. Similarly to SVR, the $\mathbf{u}_i$'s can be thought of as raw measurements or measurements lifted into a feature space, refer to Remark 5.

The classifier is obtained by solving the program:

$$\begin{array}{ll} \min\limits_{\substack{w\in\mathcal{U},b\in\mathbb{R}\\ \xi_i\geq 0,i=1,\ldots,N}} & \|w\|^2+\rho\sum\limits_{i=1}^{N}\xi_i \\ \text{subject to:} & 1-y_i(\langle w,\mathbf{u}_i\rangle-b)\leq\xi_i,\quad i=1,\ldots,N, \end{array} \tag{18}$$

which gives the classifier $\hat{y}=\text{sign}(\langle w^*,\mathbf{u}\rangle-b^*)$ ("$*$" denotes the solution to (18)).

Remark 8 *In case of lifting into a feature space, considering that the solution takes the form $w=\sum_j\alpha_j\mathbf{u}_j$ (see below), program* (18) *can be re-written as*

$$\begin{array}{ll} \min\limits_{\substack{\alpha_j,j=1,\ldots,N,b\in\mathbb{R}\\ \xi_i\geq 0,i=1,\ldots,N}} & \sum\limits_{j,k}\alpha_j\alpha_k k(\mathbf{m}_j,\mathbf{m}_k)+\rho\sum\limits_{i=1}^{N}\xi_i \\ \text{subject to:} & 1-y_i\left(\sum\limits_j\alpha_j k(\mathbf{m}_j,\mathbf{m}_i)-b\right)\leq\xi_i,\quad i=1,\ldots,N, \end{array}$$

with the classifier given by $\hat{y}=\text{sign}(\sum_j\alpha_j^* k(\mathbf{m}_j,\mathbf{m})-b^*)$. ⋆

Existence and uniqueness of the solution (w^*,b^*,ξ_i^*) present no difficulties. In contrast, non-accumulation raises some subtle issues (which refer to the situation where $w^*=0$) that make a rigorous application of the results from Section 2 non-trivial.

Existence: As in previous support vector methods, w^* must belong to a finite dimensional subspace spanned by $\{\mathbf{u}_i,i=1,\ldots,N\}$ and an optimal solution certainly exists.

Uniqueness: w^* is unique while b^* may not be, see Theorem 2 in Burges and Crisp (1999). Break the tie by minimizing $|b+1|$.[7] Similarly to SVR, this returns unique w^* and b^* and the ξ_i^*'s also remain uniquely determined, see Remark 5.

Non-accumulation: It requires satisfaction of the condition

$$\mathbb{P}\{1-y(\langle w,\mathbf{u}\rangle-b)=0\}=0\quad\forall w,b.$$

A problem with this condition rises for $w=0$ and $b=\pm 1$, in which case the condition becomes

$$\mathbb{P}\{1\pm y=0\}=0,$$

which is generally not satisfied. This is sign of an intrinsic difficulty: if one sees all labels of one type -1 or 1 (which happens with nonzero probability), then program (18) returns $w^*=0$ and $-b^*=1$ (in case of all labels equal to 1) or $-b^*=-1$ (in case of all labels equal

7. The reason for choosing $|b+1|$ and not $|b|$ is that this prevents the solution $w^*=0$ and $b^*=0$ from happening (which would result in a not-well defined classifier, see below).

to -1). Then, one ends up in a *degenerate* situation where the solution is identified by various subsets of the data set (think of when all labels are 1: any non-empty subset of data points returns the same solution), which is exactly what the non-accumulation Assumption 3 rules out. Moreover, seeing all labels of one type is not the only case in which $w^* = 0$ and $b^* = \pm 1$ and it is easy to figure out other configurations of data points for this to happen. In all these cases, degeneracy occurs. Hence, the fact that the non-accumulation Assumption 3 is not satisfied is not accidental and has deep motivations. Nevertheless, we can get around this difficulty and get the theory to work for a *heated* version of the problem. By a *cooling* procedure, one then finds rigorous results for SVM. Along this route, we also introduce a breakdown of the initial optimization problem into three distinct problems where a problem that has a specific simple structure is considered when one knows that $w^* = 0$ for the initial problem (18); this is instrumental to finding tight evaluations of the risk for this case as well. One side effect of this process is that in the final result the confidence parameter is elevated from the value β to the value 3β, which has however very little impact in practice. The technically articulated theory is presented in Appendix C, while we give here the final result. The result requires that $\mathbf{u}$ are generically distributed and do not concentrate on linear manifolds, as the following assumption states.

Assumption 6 *Assume that*

$$\mathbb{P}\{(\mathbf{u}, y) : \langle a, \mathbf{u}\rangle - h = 0\} = 0 \quad \forall a \neq 0, h.$$

$\star$

Theorem 5 (Violation of SVM) *With $\underline{\epsilon}(\cdot)$ and $\overline{\epsilon}(\cdot)$ as defined in Theorem 1, we have*

$$\mathbb{P}^N\{\underline{\epsilon}(s^*) \leq \mathbb{P}\{(\mathbf{u}, y) : 1 - y(\langle w^*, \mathbf{u}\rangle - b^*) > 0\} \leq \overline{\epsilon}(s^*)\} \geq 1 - 3\beta,$$

where s^* is so defined: when $w^* \neq 0$, s^* is the number of $(\mathbf{u}_i, y_i)$'s for which $1 - y_i(\langle w^*, \mathbf{u}_i\rangle - b^*) \geq 0$ and, when $w^* = 0$, s^* is the number of data points whose label belongs to the class with fewer elements (if, e.g., there are 960 data points with label 1 and 40 with label -1, then $s^* = 40$; if there is a fifty-fifty split, then s^* is equal to half of the data points). $\star$

Proof See Appendix C. ■

One further point that needs be clearly highlighted is that in SVM constraints violation does not correspond to misclassification. This marks a difference with SVR and SVDD where indeed constraints violation meant misprediction and was the final quantity that we wanted to keep under control. To understand this point, refer to the classifier generated by SVM:

$$\begin{aligned} &\text{classify as 1 points } \mathbf{u} \text{ such that } \langle w^*, \mathbf{u}\rangle - b^* > 0; \\ &\text{classify as } -1 \text{ points } \mathbf{u} \text{ such that } \langle w^*, \mathbf{u}\rangle - b^* < 0. \end{aligned}$$

Hence, we make an error if $(\mathbf{u}, y)$ is such that

$$y(\langle w^*, \mathbf{u}\rangle - b^*) < 0, \tag{19}$$

corresponding to having disagreement between the classifier and the actual sign of y. This condition is more restrictive than constraints violation, and in fact (19) implies (and is not implied by)

$$1 - y(\langle w^*, \mathbf{u}\rangle - b^*) > 0.$$

Hence, misclassification occurs more rarely than constraints violation. As a consequence, Theorem 5 can be used to only upper bound the probability of misclassification, a result that is stated in the next theorem.

Theorem 6 (Misclassification of SVM) *Define $\bar{\epsilon}(\cdot)$ as in Theorem 1. We have*

$$\mathbb{P}^N\{\mathbb{P}\{(\mathbf{u}, y) : y \text{ is misclassified}\} \leq \bar{\epsilon}(s^*)\} \geq 1 - 3\beta, \tag{20}$$

where s^* is so defined: when $w^* \neq 0$, s^* is the number of $(\mathbf{u}_i, y_i)$'s for which $1 - y_i(\langle w^*, \mathbf{u}_i\rangle - b^*) \geq 0$ and, when $w^* = 0$, s^* is the number of data points whose label belongs to the class with fewer elements. ⋆

Remark 9 (Sensitivity and specificity) *The probability of misclassification $\mathbb{P}\{(\mathbf{u}, y) : y \text{ is misclassified}\}$ is also known in the machine learning literature as "accuracy". Depending on the application at hand, it may be that the user is also interested in the so-called "specificity" and "sensitivity", which are the probability of misclassification within a given class, either 1 or -1 (in formulas: specificity $= \mathbb{P}\{(\mathbf{u}, y) : y \text{ is misclassified} \mid y = 1\}$; sensitivity $= \mathbb{P}\{(\mathbf{u}, y) : y \text{ is misclassified} \mid y = -1\}$). The theory that has been presented here can be easily modified to also provide a characterization of specificity and sensitivity, see Caré et al. (2018) for a similar argument applied to a different context. We sketch in this remark how this is achieved. Consider e.g. specificity. Let N_1^* be the number of observations, among the N available, for which $y_i = 1$. Take conditioning on the value of N_1^* and on the value of the remaining $N - N_1^*$ observations for which $y_i = -1$. The N_1^* observations with $y_i = 1$ are instead let vary and seen as random. With the definition that $\tilde{\mathbb{P}}_1$ is the probability over $\mathcal{U} \times \{-1, 1\}$ obtained after conditioning on $y = 1$ (that is, $\tilde{\mathbb{P}}_1(E) = \mathbb{P}(E \mid y = 1)$ for every measurable event $E \subseteq \mathcal{U} \times \{-1, 1\}$), the argument in the proof of Theorems 5 and 6 can be repeated* mutatis mutandis *to obtain*

$$\tilde{\mathbb{P}}_1^{N_1^*}\{\tilde{\mathbb{P}}_1\{(\mathbf{u}, y) : y \text{ is misclassified}\} \leq \bar{\epsilon}_{N_1^*}(s_1^*)\} \geq 1 - 3\beta, \tag{21}$$

where $\bar{\epsilon}_{N_1^*}$ is the same function as $\bar{\epsilon}$ (calculated for N_1^* observations) and s_1^* is defined similarly to s^* in Theorem 5 but limited to observations for which $y_i = 1$ (in the proof, observations $y_i = -1$ have to be thought of as fixed and they do not concur in the evaluation of the complexity). Next, integrating the result in (21) *over the value of N_1^* and over the variability of the observations with $y_i = -1$ yields*

$$\mathbb{P}^N\{\mathbb{P}\{(\mathbf{u}, y) : y \text{ is misclassified} \mid y = 1\} \leq \bar{\epsilon}_{N_1^*}(s_1^*)\} \geq 1 - 3\beta,$$

which is the characterization of specificity. Sensitivity is dealt with analogously. ⋆

4. Numerical Example

Inspired by the numerical example in Schölkopf et al. (1998), we applied SVR to find a regression model for points generated by a noisy sinc function. Specifically, we considered a data set formed by $N = 2000$ examples $(\mathbf{m}_i, y_i)$ with $\mathbf{m}_i$ extracted uniformly from $[-3, 3]$ and $y_i = \sin(\pi \mathbf{m}_i)/(\pi \mathbf{m}_i) + e_i$, where e_i had a Laplace distribution with mean $\mu = 0$ and parameter $b = 1$ (data points are in Figure 5). We used the Gaussian kernel $\exp(-|\mathbf{m}_k - \mathbf{m}_j|^2/\sigma^2)$, where σ was regarded as an adjustable hyper-parameter.[8] With $\tau = 0.01$, program (13) was repeatedly solved with $\sigma = 10^k$, $k = -2, -1, 0, 1$ and $\rho = (3/5)^\ell$, $\ell = 0, 1, \ldots, 14$. Each time, we recorded: the solution; the cost; the value of the complexity s^*; and the interval for the risk $[\underline{\epsilon}(s^*), \overline{\epsilon}(s^*)]$ with $\beta = 10^{-4}$, which was calculated as indicated in Theorem 3.

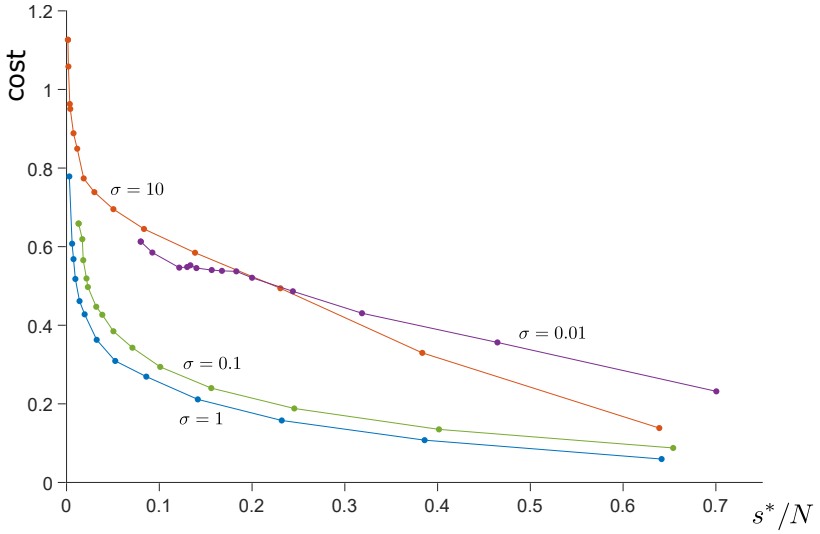

Figure 4: cost vs. value of s^*/N for the various solutions.

Figure 4 shows the results: the y-axis gives the value of the cost while the x-axis contains the value of s^*/N (which is an indicator of the risk). From this figure, it appears that $\sigma = 1$ dominates over other choices of σ.

Focusing on the solutions obtained for $\sigma = 1$ we then constructed a cost-risk plot putting on display the cost and the corresponding interval for the risk ($\beta = 10^{-4}$) obtained for various values of ρ. The plot is that of Figure 2. For $\rho = 1$ we obtained the smallest value for s^*, whence the range for the risk hit its minimum, at the expense of a large cost. As ρ was decreased, s^* showed a monotonic growth. Initially, the drop in the cost was rapid, paired with a moderate increase of the risk. Instead, for smaller values of ρ, even a small decrease of cost implied a significant rise of the risk. We opted for $\bar{\rho} = (3/5)^9$, yielding $s^* = 105$ (corresponding to $[\underline{\epsilon}(105), \overline{\epsilon}(105)] = [0.032, 0.08]$) and $c(x^*) = 0.31$. Since τ was

8. Note that by means of parameter σ one can tune the locality of the basis functions associated to the Gaussian kernel and selecting a small σ corresponds to fine-grained basis functions with better descriptive capabilities (tantamount to what is achieved with polynomial kernels by increasing their order). Thus, small σ's correspond to improved cost values, but also to increased solution complexities and, therefore, to higher risks.

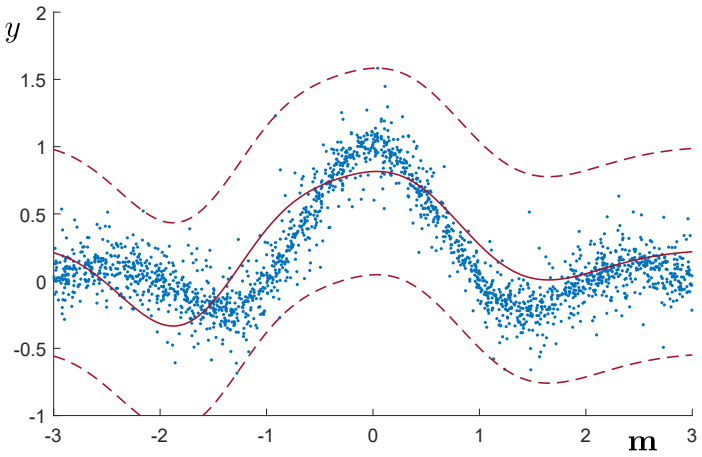

Figure 5: SVR model for $\rho = 1$.

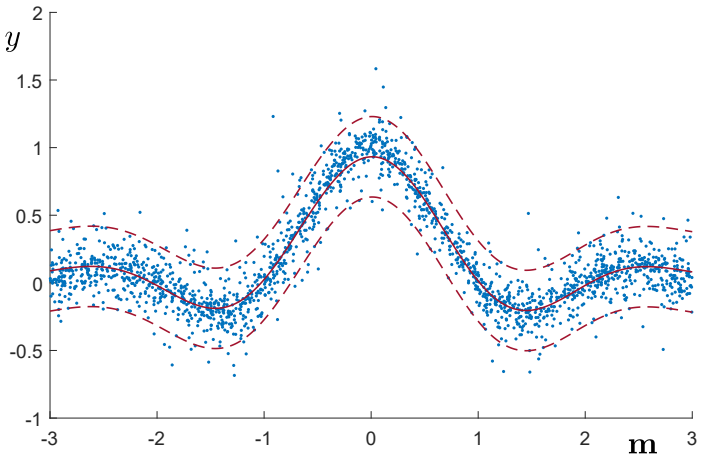

Figure 6: SVR model for $\rho = (3/5)^9$.

small, the cost $c(x^*)$ is almost identical to γ^*, the size of the tube. Figures 5, 6, and 7 depict the models obtained for $\rho = 1$, $\rho = (3/5)^9$ (our choice), and $\rho = (3/5)^{14}$. A visual inspection, possible in this case because we are considering a toy example with $\mathbf{m}$ scalar, confirms the analysis based on the ground of the cost-risk plot.

Finally, we thought it might be useful to look closer at the validity and tightness of Theorem 1. While keeping ρ at the value $(3/5)^9$, we solved problem (13) 200 times, each time drawing a new sample of size 2000. Each solution was then tested on 10000 additional random values of $(\mathbf{m}, y)$ to evaluate its risk (Monte-Carlo approach). Figure 8 plots the pairs (complexity,risk) obtained in the 200 trials, along with the upper and lower limits $\underline{\epsilon}(k)$ and $\overline{\epsilon}(k)$ for $\beta = 10^{-4}$. Theorem 1 predicts that the risk is, on average, in the interval $\underline{\epsilon}(k)$ and $\overline{\epsilon}(k)$ in 9999 times out of 10000. This was the case for all the 200 points in our simulations. A visual inspection also reveals that the spread of the risks fills well the vertical range given by the theoretical bounds, a sign that the theoretical result provides tight evaluations in spite of its prerogative of being distribution-free.

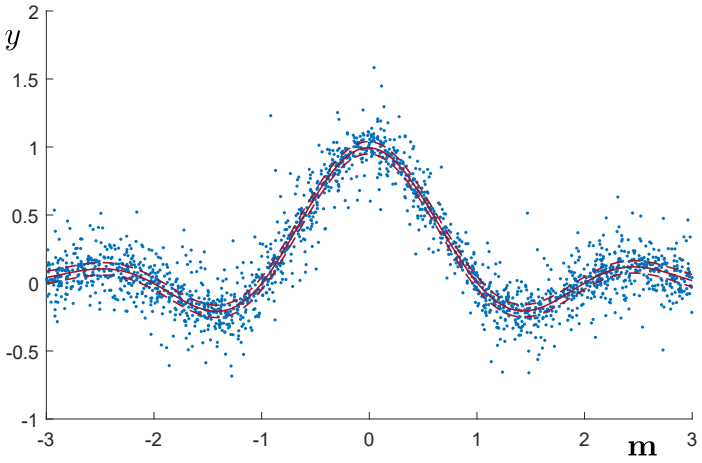

Figure 7: SVR model for $\rho = (3/5)^{14}$.

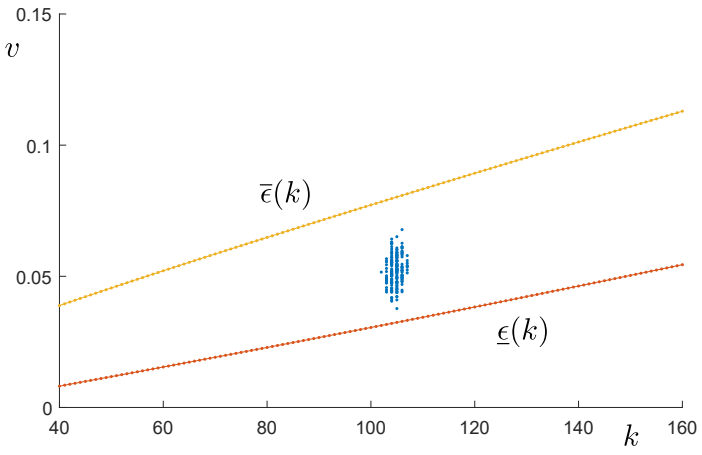

Figure 8: (complexity,risks) pairs (blue dots) vs. $\underline{\epsilon}(k)$ and $\overline{\epsilon}(k)$ (continuous dotted lines); $N = 2000$ and $\beta = 10^{-4}$.

Acknowledgments

The authors are indebtedd to Nicolò Cesa-Bianchi and Steve Hanneke for insightful comments and discussion on the content of this paper.

Appendix A. Proof of Theorem 2

Let $v := 1 - t$. Equation (5) for $k = 0, \ldots, N-1$ becomes

$$\frac{\beta}{2N}\sum_{i=k}^{N-1}\binom{i}{k}(1-v)^{i-k} + \frac{\beta}{6N}\sum_{i=N+1}^{4N}\binom{i}{k}(1-v)^{i-k} = \binom{N}{k}(1-v)^{N-k}. \tag{22}$$

The fact that (5) has two solutions in $[0, +\infty)$, as stated in Theorem 1, translates into that equation (22) has two solutions in $(-\infty, 1]$, namely $\underline{\epsilon}(k)$ and $\overline{\epsilon}(k)$. Observing that the left-hand side of (22) is equal to $\beta/2N > 0$ for $v = 1$, while the right-hand side is zero at the same point, we then conclude that, when running backward from 1 to $-\infty$, the left-hand side is first above, then below, and then above again of the right-hand side, as graphically

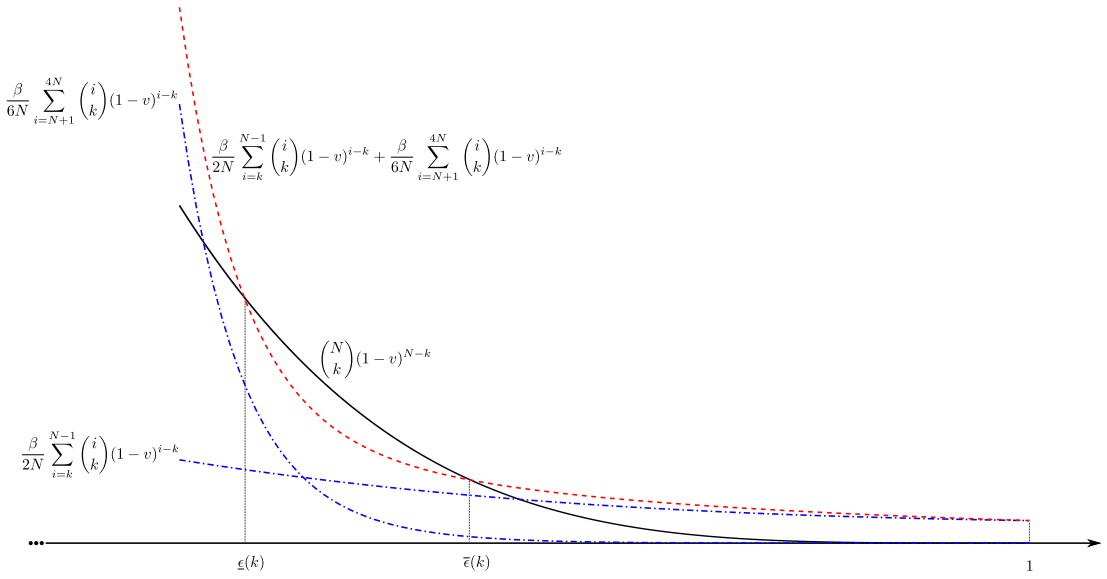

Figure 9: Graphical representation of functions in proof of Theorem 2.

illustrated in Figure 9.
Next consider the following two inequalities

$$\frac{\beta}{2N}\sum_{i=k}^{N-1}\binom{i}{k}(1-v)^{i-k} \geq \binom{N}{k}(1-v)^{N-k}, \tag{23}$$

$$\frac{\beta}{6N}\sum_{i=N+1}^{4N}\binom{i}{k}(1-v)^{i-k} \geq \binom{N}{k}(1-v)^{N-k}. \tag{24}$$

These two inequalities can be used to effectively locate a suitable upper-bound for $\overline{\epsilon}(k)$ (inequality (23)) and lower-bound for $\underline{\epsilon}(k)$ (inequality (24)). This is explained as follows. Take the ratio of the left-hand side over the right-hand side of equation (23):

$$\frac{\beta}{2N}\sum_{i=k}^{N-1}\frac{\binom{i}{k}}{\binom{N}{k}}(1-v)^{i-N}.$$

Over $(-\infty,1)$, this function is strictly increasing, moreover for $v=0$ it is smaller than $\beta/2<1$ (note that $\frac{\binom{i}{k}}{\binom{N}{k}}<1$) while it tends to $+\infty$ as $v\to 1$. Therefore, it picks the value 1 in one and only one point in $(0,1)$, which shows that equality is attained in (23) for only one value of $v\in(0,1)$. Hence, the two functions showing up in the left-hand and right-hand sides of (23) are mutually positioned as shown in Figure 9.
Further, it is claimed that any v satisfying (23) is an upper-bound to $\overline{\epsilon}(k)$. Indeed, when moving from equation (22) to (23) we have removed from the left-hand side of (22) a positive term, so shifting to the right the point where equality is achieved in (23); then, owing to the mutual position of the two functions in (23) one immediately sees the correctness of the claim.
The inequality condition (24) can be studied in full analogy to (23) with the only advisory that the role of interval $(0,1)$ is played by $(-\infty,1)$ when considering the second inequality (24).

Preliminary calculations

To study (23) and (24), we shall use a re-writing of the right-hand sides of these inequalities as given in the following.
Let

$$\varphi_{H,k}(v)=\sum_{i=k}^{H-1}\binom{i}{k}(1-v)^{i-k}.$$

Notice first that, for $k=0$, we have $\varphi_{H,0}(v)=\sum_{i=0}^{H-1}(1-v)^i=\frac{1-(1-v)^H}{v}$. Next, for $k\leq H-1$, a direct verification proves the validity of the following updating rule

$$\varphi_{H,k}(v)=-\frac{1}{k}\frac{\mathrm{d}}{\mathrm{d}v}\varphi_{H,k-1}(v). \tag{25}$$

A repeated use (a cumbersome but straightforward exercise) of (25) now gives

$$\varphi_{H,k}(v) = \frac{1-\sum_{i=0}^{k}\binom{H}{i}v^i(1-v)^{H-i}}{v^{k+1}} \tag{26}$$

$$= \frac{\sum_{i=k+1}^{H}\binom{H}{i}v^i(1-v)^{H-i}}{v^{k+1}}. \tag{27}$$

Upper bounding $\overline{\epsilon}(k)$

Substituting (26) in (23), (23) becomes

$$\frac{\beta}{2}\left(1-\sum_{i=0}^{k}\binom{N}{i}v^i(1-v)^{N-i}\right)\geq N\binom{N}{k}v^{k+1}(1-v)^{N-k}. \tag{28}$$

If we further decrease the left-hand side (and increase the right-hand side) we obtain an inequality the solutions of which are still upper-bounds to $\bar{\epsilon}(k)$. Starting with the left-hand side, we apply an argument first used in Alamo et al. (2015) and, for any $a > 1$, write:

$$\begin{aligned}
&\sum_{i=0}^{k} \binom{N}{i} v^i (1-v)^{N-i} \\
&\quad \leq \quad a^k \sum_{i=0}^{k} \binom{N}{i} \left(\frac{v}{a}\right)^i (1-v)^{N-i} \\
&\quad \leq \quad a^k \sum_{i=0}^{N} \binom{N}{i} \left(\frac{v}{a}\right)^i (1-v)^{N-i} \\
&\quad = \quad a^k \left(1 - v + \frac{v}{a}\right)^N \\
&\quad = \quad (1-(1-a))^k \left(1 - \frac{a-1}{a} v\right)^N \\
&\quad \leq \quad \mathrm{e}^{-(1-a)k} \mathrm{e}^{-\frac{a-1}{a} vN}, \qquad (29)
\end{aligned}$$

where the last inequality follows from relation $1 - z \leq \mathrm{e}^{-z}$. Similarly, using also the fact that $N\binom{N}{k} \leq (k+1)\binom{N+1}{k+1}$,

$$\begin{aligned}
&N \binom{N}{k} v^{k+1} (1-v)^{N-k} \\
&\quad \leq \quad (k+1) \binom{N+1}{k+1} v^{k+1} (1-v)^{N+1-(k+1)} \\
&\quad \leq \quad (k+1) \sum_{i=0}^{k+1} \binom{N+1}{i} v^i (1-v)^{N+1-i} \\
&\quad \leq \quad (k+1) \mathrm{e}^{-(1-a)(k+1)} \mathrm{e}^{-\frac{a-1}{a} v(N+1)} \\
&\quad \leq \quad (k+1) \mathrm{e}^{-(1-a)} \mathrm{e}^{-(1-a)k} \mathrm{e}^{-\frac{a-1}{a} vN}. \qquad (30)
\end{aligned}$$

Suppose now $k > 0$ (the case $k = 0$ will be considered separately) and take $a = 1 + 1/\sqrt{k}$. Using (29) and (30) in (28) yields that any v coming from the inequality

$$\frac{\beta}{2} \left(1 - \mathrm{e}^{\sqrt{k}} \mathrm{e}^{-\frac{vN}{\sqrt{k}+1}}\right) \geq (k+1) \mathrm{e}^{\frac{1}{\sqrt{k}}} \mathrm{e}^{\sqrt{k}} \mathrm{e}^{-\frac{vN}{\sqrt{k}+1}}$$

is an upper bound to $\bar{\epsilon}(k)$. This inequality is equivalent to

$$\frac{\beta}{2(k+1)} \geq \mathrm{e}^{\sqrt{k}} \mathrm{e}^{-\frac{vN}{\sqrt{k}+1}} \left[\frac{\beta}{2(k+1)} + \mathrm{e}^{\frac{1}{\sqrt{k}}}\right]$$

and, solving for v, we obtain

$$v \geq \frac{k}{N} + \frac{\sqrt{k}+1}{N} \left(\lambda + \ln \frac{2}{\beta} + \ln(k+1)\right),$$

where $\lambda = \ln\left[\frac{\beta}{2(k+1)} + \mathrm{e}^{\frac{1}{\sqrt{k}}}\right] + \frac{\sqrt{k}}{\sqrt{k}+1}$. This shows that

$$\overline{\epsilon}(k) \leq \frac{k}{N} + \frac{\sqrt{k}+1}{N}\left(\lambda + \ln\frac{2}{\beta} + \ln(k+1)\right)$$

and the validity of (11) (for $k \neq 0, N$ – recall that we started from equation (5) that holds for $k < N$ and further left behind the case k=0) follows by noticing that $\lambda \leq 2$.
Turn now to the remaining cases, $k = 0$ or $k = N$.
Case $k = N$ is trivial because $\overline{\epsilon}(N) = 1$, which is clearly in agreement with (11).
As for $k = 0$, go back to (28) and use in it (29) and (30) with $a = 1 + 1/\sqrt{k+1}$, which, after substituting $k = 0$, gives $a = 2$ (adding 1 to k serves the purpose of avoiding division by zero). Operating the same manipulations as before we now obtain

$$v \geq \frac{2}{N}\left(\ln\left[\frac{\beta}{2} + \mathrm{e}\right] + \ln\frac{2}{\beta}\right),$$

which has the form of the upper bound for $\overline{\epsilon}(k)$ given in Theorem 2.

Lower bounding $\underline{\epsilon}(k)$

First, we want to claim that for any k large enough there is a positive v satisfying equation (24). In fact, for $v = 0$ equation (24) reduces to $\frac{\beta}{6N}\sum_{i=N+1}^{4N}\binom{i}{k} \geq \binom{N}{k}$ and, using the hockey-stick identity (i.e., $\sum_{i=r}^{n}\binom{i}{r} = \binom{n+1}{r+1}$), we have

$$\begin{aligned}
&\frac{\beta}{6N}\frac{\sum_{i=N+1}^{4N}\binom{i}{k}}{\binom{N}{k}} \\
&= \frac{\beta}{6N}\frac{\binom{4N+1}{k+1} - \binom{N+1}{k+1}}{\binom{N}{k}} \\
&= \frac{\beta}{6N}\frac{(4N+1)\cdots(4N-k+1) - (N+1)\cdots(N-k+1)}{(N)\cdots(N-k+1)\cdot(k+1)} \\
&\geq \frac{\beta}{6}\frac{2^{k+1}(N+1)\cdots(N-k+1) - (N+1)\cdots(N-k+1)}{(N+1)\cdots(N-k+1)\cdot(k+1)} \\
&= \frac{\beta}{6}\frac{2^{k+1}-1}{k+1},
\end{aligned}$$

which is greater than 1 for any

$$k \geq c_1 + c_2\ln(1/\beta), \tag{31}$$

where c_1 and c_2 are suitable constants. In what follows, we assume that this latter condition is satisfied and hence seek a positive solution of equation (24).
The summation in the left-hand side of (24) can be re-written as $\sum_{i=N+1}^{4N}\binom{i}{k}(1-v)^{i-k} =$

$\varphi_{4N+1,k}(v) - \varphi_{N+1,k}(v)$, which, owing to (27), allows us to re-write equation (24) as follows

$$\frac{\beta}{6}\left(\sum_{i=k+1}^{4N+1}\binom{4N+1}{i}v^i(1-v)^{4N+1-i} - \sum_{i=k+1}^{N+1}\binom{N+1}{i}v^i(1-v)^{N+1-i}\right) \geq N\binom{N}{k}v^{k+1}(1-v)^{N-k}, \tag{32}$$

where moving term v^{k+1} to the right-hand side does not change the inequality sign because v is positive. Similarly to what we did to find an upper bound for $\overline{\epsilon}(k)$, here we can decrease the left-hand side and increase the right-hand side of (32) to find a valid lower bound for $\underline{\epsilon}(k)$.
Notice first that $\sum_{i=k+1}^{4N+1}\binom{4N+1}{i}v^i(1-v)^{4N+1-i} \geq \frac{1}{2}$ for $v \geq \frac{k+1}{4N+2}$.[9] Thus, using again the fact $N\binom{N}{k} \leq (k+1)\binom{N+1}{k+1}$, we can take

$$\frac{\beta}{6}\left(\frac{1}{2} - \sum_{i=k+1}^{N+1}\binom{N+1}{i}v^i(1-v)^{N+1-i}\right) \geq (k+1)\binom{N+1}{k+1}v^{k+1}(1-v)^{N+1-(k+1)} \tag{33}$$

in place of (32) to obtain a lower bound to $\underline{\epsilon}(k)$ as long as we impose the additional condition that

$$v \geq \frac{k+1}{4N+2}. \tag{34}$$

For any $a > 1$, we now have

$$\begin{aligned}
&\binom{N+1}{k+1}v^{k+1}(1-v)^{N+1-(k+1)} \\
&\leq \sum_{i=k+1}^{N+1}\binom{N+1}{i}v^i(1-v)^{N+1-i} \\
&\leq \frac{1}{a^k}\sum_{i=k+1}^{N+1}\binom{N+1}{i}(av)^i(1-v)^{N+1-i} \\
&\leq \frac{1}{a^k}\sum_{i=0}^{N+1}\binom{N+1}{i}(av)^i(1-v)^{N+1-i} \\
&= \frac{1}{a^k}\left(1+(a-1)v\right)^{N+1} \\
&\leq \frac{\mathrm{e}^{(a-1)v(N+1)}}{a^k},
\end{aligned}$$

where the last inequality follows from relation $1 + z \leq \mathrm{e}^z$. Assume $k > 0$ and take $a = 1 + 1/\sqrt{k}$. Using the above chain of inequalities twice in (33) (for the term in the left-hand side of (33) we use the inequality obtained by comparing the second with the last term in

9. This follows from the fact that $\sum_{i=k+1}^{4N+1}\binom{4N+1}{i}v^i(1-v)^{4N+1-i}$ is the cumulative distribution function of a Beta distribution and $\frac{k+1}{4N+2}$ is its mean, which is greater than the median, Payton et al. (1989).

the chain), we obtain the following condition that is more restrictive than (33)

$$\frac{\beta}{6}\left(\frac{1}{2}-\frac{\mathrm{e}^{\frac{v(N+1)}{\sqrt{k}}}}{(1+\frac{1}{\sqrt{k}})^k}\right)\geq(k+1)\frac{\mathrm{e}^{\frac{v(N+1)}{\sqrt{k}}}}{(1+\frac{1}{\sqrt{k}})^k}.$$

This inequality is equivalent to

$$\frac{\beta}{12(\frac{\beta}{6}+k+1)}\geq\frac{\mathrm{e}^{\frac{v(N+1)}{\sqrt{k}}}}{(1+\frac{1}{\sqrt{k}})^k},$$

which, solved for v, gives

$$v\leq\frac{k}{N+1}\ln\left[\left(1+\frac{1}{\sqrt{k}}\right)^{\sqrt{k}}\right]-\frac{\sqrt{k}}{N+1}\left(\ln\frac{12}{\beta}+\ln\left(\frac{\beta}{6}+k+1\right)\right).$$

Noticing now that $\ln(1+x)\geq x-x^2/2$ for all $x\geq 0$, we can finally replace the latter inequality with

$$v\leq\frac{k}{N+1}\left(1-\frac{1}{2\sqrt{k}}\right)-\frac{\sqrt{k}}{N+1}\left(\ln\frac{12}{\beta}+\ln\left(\frac{\beta}{6}+k+1\right)\right), \tag{35}$$

which, for a more handy use, we also rewrite as

$$v\leq\frac{k}{N}-g(k,N,\beta),$$

where function $g(k,N,\beta)$ is just the difference between k/N and the right-hand side of (35). Notice also that this equation is valid also for $k=N$ since (6) also leads to (24), which has been our starting point in the derivation.

To conclude the proof, we have to put together all inequalities that limit the choice of v, namely:

(i) $k\geq c_1+c_2\ln(1/\beta)$ (equation (31));

(ii) $v\geq\frac{k+1}{4N+2}$ (equation (34));

(iii) $v\leq\frac{k}{N}-g(k,N,\beta)$.

Recall that (iii) makes sense only for $k\neq 0$, however this is of no concern because the case $k=0$ takes care of itself since Theorem 2 claims that $\underline{\epsilon}(0)\geq 0$ which is in agreement with the value of $\underline{\epsilon}(0)$ given in Theorem 1. For the time being, leave (i) behind. Now, one can take the value of v that achieves equality in (iii), i.e., $v=\frac{k}{N}-g(k,N,\beta)$, provided that this is compatible with (ii), that is, $\frac{k}{N}-g(k,N,\beta)\geq\frac{k+1}{4N+2}$. This can be re-written as $g(k,N,\beta)\leq\frac{k}{N}-\frac{k+1}{4N+2}$. Instead, for those values of k,N,β for which this latter inequality does not hold, we have $g(k,N,\beta)>\frac{k}{N}-\frac{k+1}{4N+2}$, from which an easy calculation shows that $2g(k,N,\beta)\geq\frac{k}{N}$, or, equivalently, $\frac{k}{N}-2g(k,N,\beta)\leq 0$. Since $\underline{\epsilon}(k)\geq 0$, we conclude that in

any case $\underline{\epsilon}(k) \geq \frac{k}{N} - 2g(k,N,\beta)$. Noticing now that $g(k,N,\beta)$ can be upper bounded by $C'\frac{\sqrt{k}\ln\frac{1}{\beta}+\sqrt{k}\ln k+1}{N}$ for a suitable value of the constant C', we conclude that

$$\underline{\epsilon}(k) \geq \frac{k}{N} - C\frac{\sqrt{k}\ln\frac{1}{\beta}+\sqrt{k}\ln k+1}{N}, \quad \text{where } C = 2C'. \tag{36}$$

Turn now back to consider (i). Condition (i) is not satisfied when $\frac{k}{N} < (c_1 + c_2\ln(1/\beta))/N$. However, this latter condition implies that the right-hand side of (36) is negative (possibly after enlarging the constant C in (36) to a value that, with a little abuse of notation, we still call C), so that (36) is always a valid lower bound because $\underline{\epsilon}(k)$ is always non-negative. This concludes the proof. ■

Appendix B. Proof that $\tilde{\epsilon}(c,0) \geq 1-(1-c/N)(c/N)^{\frac{c/N}{1-c/N}}$ Asymptotically

The proof is based on the following bounds to the factorial of an integer, Robbins (1955):

$$\sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi}n^{n+\frac{1}{2}}e^{-n}e^{\frac{1}{12n}}. \tag{37}$$

Start by noticing that

$$\begin{aligned}
\tilde{\epsilon}(c,0) &= -\ln\left[\left(\frac{\beta}{N\binom{N}{c}}\right)^{\frac{1}{N-c}}\right] \\
&\geq 1-\left(\frac{\beta}{N\binom{N}{c}}\right)^{\frac{1}{N-c}} \\
&\geq 1-\left(\frac{\beta}{\binom{N}{c}}\right)^{\frac{1}{N-c}} \\
&= 1-\left(\frac{\beta c!(N-c)!}{N!}\right)^{\frac{1}{N-c}},
\end{aligned}$$

where the first inequality holds because $-\ln(x) \geq 1-x$. Using (37) to bound the factorials at the numerator and denominator in the last expression now yields

$$\begin{aligned}
\tilde{\epsilon}(c,0) &\geq 1-\left(\frac{\beta\sqrt{2\pi}c^{c+\frac{1}{2}}e^{-c}e^{\frac{1}{12c}}\sqrt{2\pi}(N-c)^{N-c+\frac{1}{2}}e^{-N+c}e^{\frac{1}{12(N-c)}}}{\sqrt{2\pi}N^{N+\frac{1}{2}}e^{-N}e^{\frac{1}{12N+1}}}\right)^{\frac{1}{N-c}} \\
&= 1-\left(\beta\sqrt{2\pi}\right)^{\frac{1}{N-c}} \times e^{\left(\frac{1}{12c}+\frac{1}{12(N-c)}-\frac{1}{12N+1}\right)\frac{1}{N-c}} \times \left(\frac{c(N-c)}{N}\right)^{\frac{1}{2(N-c)}} \times \\
&\qquad \times \left(\frac{c^c(N-c)^{N-c}}{N^N}\right)^{\frac{1}{N-c}} \\
&= 1-\left(\beta\sqrt{2\pi}\right)^{\frac{1}{N-c}} \times e^{\left(\frac{1}{12c}+\frac{1}{12(N-c)}-\frac{1}{12N+1}\right)\frac{1}{N-c}} \times \left(\frac{c(N-c)}{N}\right)^{\frac{1}{2(N-c)}} \times \\
&\qquad \times \left(1-\frac{c}{N}\right)\left(\frac{c}{N}\right)^{\frac{\frac{c}{N}}{1-\frac{c}{N}}}.
\end{aligned}$$

Take now $c = \mu N$. The first three terms in the product in the last expression tend to 1 as $N \to \infty$. Whence,

$$\tilde{\epsilon}(d,0) \geq 1 - \left(1 - \frac{c}{N}\right)\left(\frac{c}{N}\right)^{\frac{\frac{c}{N}}{1-\frac{c}{N}}}$$

as $N \to \infty$. This concludes the proof. ■

Appendix C. Proof of Theorem 5

For analysis purposes, introduce the augmented probability space $(\mathcal{U} \times \{-1,1\}) \times [0,1]$ endowed with the probability $\mathbb{Q} = \mathbb{P} \times \mathbb{U}$, where $\mathbb{U}$ is the uniform probability on $[0,1]$ that describes the "heating variable" z. Next, fix a real parameter value α chosen from the countable set $\{1/j\}$, where j is any positive integer, and consider an independent heated data set $\{\mathbf{u}_i, y_i, (1-\alpha z_i)\}_{i=1}^N$ generated from $((\mathcal{U}\times\{-1,1\})\times[0,1])^N$. Note that this situation traces back to the actual data generation mechanism when $\alpha \to 0$ because variable z loses its heating role and augmenting $(\mathcal{U} \times \{-1,1\})$ with $[0,1]$ has no effect.

Suppose we run program (18) with the heated data set, that is, we run

$$\begin{array}{ll} \min\limits_{\substack{w\in\mathcal{U}, b\in\mathbb{R} \\ \xi_i\geq 0, i=1,\ldots,N}} & \|w\|^2 + \rho\sum\limits_{i=1}^N \xi_i \\ \text{subject to:} & (1-\alpha z_i) - y_i(\langle w, u_i\rangle - b) \leq \xi_i, \quad i = 1,\ldots,N, \end{array} \tag{38}$$

endowed with the same rule adopted in (18) to break the tie in case of non-unique solution. Then, existence and uniqueness are preserved and it is further claimed that the non-accumulation Assumption 3 also holds. Indeed, with heated values y, the non-accumulation condition writes $\mathbb{Q}\{(1-\alpha z) - y(\langle w, \mathbf{u}\rangle - b) = 0\} = 0$, $\forall (w,b) \in \mathcal{U} \times \mathbb{R}$, a condition that is proven by the following calculation:

$$\begin{array}{rcl} \mathbb{Q}\{(1-\alpha z) - y(\langle w, \mathbf{u}\rangle - b) = 0\} & & \\ = & \mathbb{Q}\left\{z = \dfrac{1 - y(\langle w, \mathbf{u}\rangle - b)}{\alpha}\right\} & \\ = & \mathbb{Q}\left\{\mathbb{Q}\left\{z = \dfrac{1 - y(\langle w, \mathbf{u}\rangle - b)}{\alpha} \;\middle|\; \mathbf{u}, y\right\}\right\} & \\ = & 0. & \end{array} \tag{39}$$

Hence, the result in Theorem 1 can be applied to the heated situation yielding:

$$\mathbb{Q}^N\{\underline{\epsilon}(s_\alpha^*) \leq V_\alpha(w_\alpha^*, b_\alpha^*) \leq \overline{\epsilon}(s_\alpha^*)\} \geq 1 - \beta, \tag{40}$$

where subscript α indicates that the solution has been obtained from the heated program (38), $V_\alpha(w,b) = \mathbb{Q}\{(\mathbf{u},y,z) : (1-\alpha z) - y(\langle w, \mathbf{u}\rangle - b) > 0\}$ and s_α^* is the number of $(\mathbf{u}_i, y_i, z_i)$'s for which $(1-\alpha z_i) - y_i(\langle w_\alpha^*, \mathbf{u}_i\rangle - b_\alpha^*) \geq 0$.

To re-approach the result (40) that holds for the heated situation with the initial non-heated problem, let us start by introducing the notation $V_0(w,b) := \mathbb{Q}\{(\mathbf{u},y,z) : 1 - y(\langle w, \mathbf{u}\rangle - b) > 0\}$ and note that $V(w,b) := \mathbb{P}\{(\mathbf{u},y) : 1 - y(\langle w, \mathbf{u}\rangle - b) > 0\} = V_0(w,b)$. For a given $\alpha > 0$, write

$$V_0(w^*,b^*) = (V_0(w^*,b^*) - V_\alpha(w^*,b^*)) + (V_\alpha(w^*,b^*) - V_\alpha(w_\alpha^*,b_\alpha^*)) + V_\alpha(w_\alpha^*,b_\alpha^*). \tag{41}$$

It is claimed that the first two terms in the right-hand side exhibit the following behaviour:

(i) for all realizations of $\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N$, it holds that $\lim_{\alpha\to 0}(V_0(w^*, b^*) - V_\alpha(w^*, b^*)) = 0$;

(ii) for all realizations of $\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N$ such that $w^* \neq 0$, it holds that $\lim_{\alpha\to 0}(V_\alpha(w^*, b^*) - V_\alpha(w^*_\alpha, b^*_\alpha)) = 0$.

Proof of (i): Note that w^* and b^* only depend on the training sequence and are treated as deterministic in the calculations that follow to compute the risks. Let $B_\alpha := \{(\mathbf{u}, y, z) : (1-\alpha z) - y(\langle w^*, \mathbf{u}\rangle - b^*) > 0\}$ and $B_0 := \{(\mathbf{u}, y, z) : 1 - y(\langle w^*, \mathbf{u}\rangle - b^*) > 0\}$. By a direct inspection one can show that $B_{\alpha_1} \subseteq B_{\alpha_2}$ for $\alpha_2 \leq \alpha_1$ and that $B_0 = \cup_\alpha B_\alpha$. Hence, by σ-additivity, $V_0(w^*, b^*) = \mathbb{Q}\{B_0\} = \lim_{\alpha\to 0} \mathbb{Q}\{B_\alpha\} = \lim_{\alpha\to 0} V_\alpha(w^*, b^*)$, and claim (i) remains proven.

Proof of (ii): Note that $w^*_\alpha \to w^*$ and that $b^*_\alpha \to b^*$ as $\alpha \to 0$. Moreover, by assumption $w^* \neq 0$. Let $B^\alpha_\alpha := \{(\mathbf{u}, y, z) : (1-\alpha z) - y(\langle w^*_\alpha, \mathbf{u}\rangle - b^*_\alpha) > 0\}$. Over the complement of set $A := \{(\mathbf{u}, y, z) : 1 - y(\langle w^*, \mathbf{u}\rangle - b^*) = 0\}$, for any given $(\mathbf{u}, y, z)$, the two left-hand sides in the inequalities that define B_α and B^α_α agree in sign in the limit when $\alpha \to 0$, so that, in the limit, $B_\alpha \triangle B^\alpha_\alpha \subseteq A$ ($\triangle$ denotes symmetric difference). More formally, this means that for all $(\mathbf{u}, y, z) \in A^c$, the complement of A, there exists an $\bar\alpha$ such that $(\mathbf{u}, y, z) \notin B_\alpha \triangle B^\alpha_\alpha$ for all $\alpha \leq \bar\alpha$. This property in turn implies that $\limsup_{\alpha\to 0} \mathbb{Q}\{B_\alpha \triangle B^\alpha_\alpha\} \leq \mathbb{Q}\{A\}$ and therefore we have:

$$
\begin{aligned}
&\limsup_{\alpha\to 0} |V_\alpha(w^*, b^*) - V_\alpha(w^*_\alpha, b^*_\alpha)| \\
&\quad = \limsup_{\alpha\to 0} |\mathbb{Q}\{B_\alpha\} - \mathbb{Q}\{B^\alpha_\alpha\}| \\
&\quad \leq \limsup_{\alpha\to 0} \mathbb{Q}\{B_\alpha \triangle B^\alpha_\alpha\} \\
&\quad \leq \mathbb{Q}\{A\} \\
&\quad = 0 \quad \text{(recall that } w^* \neq 0 \text{ and use Assumption 6).}
\end{aligned}
$$

This completes the proof of (ii).[10]

Using (i) and (ii) in (41), we obtain:

$$
\left.\begin{array}{c}\text{for all realizations of } \{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N \text{ such that } w^* \neq 0, \text{ it holds that} \\ \lim_{\alpha\to 0} V_\alpha(w^*_\alpha, b^*_\alpha) = V_0(w^*, b^*).\end{array}\right\} \tag{42}
$$

Turn now to consider s^* and s^*_α. We show that:

$$
\left.\begin{array}{c}\text{with the exception of a zero-probability set, for all realizations of} \\ \{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N \text{ such that } w^* \neq 0, \text{ it holds that} \\ \lim_{\alpha\to 0} s^*_\alpha = s^*.\end{array}\right\} \tag{43}
$$

10. Note that Assumption 6 cannot be dispensed for as shown by the following counterexample. Suppose that $u \in \mathbb{R}$ has mass concentrated over ± 1 with equal probability 0.5 and $y = u$. Clearly, Assumption 6 is not satisfied in this case. When the u_i are not picked all equal and ρ is large, we have $w^* = 1$ and $b^* = 0$, and $V_\alpha(w^*, b^*)$ is zero. However, with the exception of a zero-probability set, we have $w^*_\alpha \cdot 1 - b^*_\alpha < 1$ and $w^*_\alpha \cdot (-1) - b^*_\alpha > -1$, so that $V_\alpha(w^*_\alpha, b^*_\alpha) \neq 0$ with a value that depends on the realization of $\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N$, but that is constant with α. Hence, $\lim_{\alpha\to 0}(V_\alpha(w^*, b^*) - V_\alpha(w^*_\alpha, b^*_\alpha)) \neq 0$.

To see this, note that, when $w^* \neq 0$ and with the exception of a zero-probability set, Assumption 6 implies that the $(\mathbf{u}_i, y_i, z_i)$ such that $1 - y_i(\langle w^*, \mathbf{u}_i \rangle - b^*) \geq 0$ correspond to the active constraints for (18), and all of these active constraints are strictly needed to determine the solution w^*, b^*, ξ_i^*. A small enough heating keeps these and only these constraints active for (38) too, which implies that $s_\alpha^* = s^*$ for all α small enough.

Using (40), (42), and (43), we are now ready to establish results that quantify the violation when $w^* \neq 0$.

Let $I(k) = [\underline{\epsilon}(k), \overline{\epsilon}(k)]$ and define the following events in $((\mathcal{U} \times \{-1, 1\}) \times [0, 1])^N$:

$$\begin{aligned} E &= \left\{\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N : \ w^* \neq 0 \ \wedge V(w^*, b^*) \notin I(s^*)\right\} \\ E_\alpha &= \left\{\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N : \ w^* \neq 0 \ \wedge V_\alpha(w_\alpha^*, b_\alpha^*) \notin I(s_\alpha^*)\right\} \\ E_\alpha^+ &= \cap_{\alpha' \leq \alpha} E_{\alpha'}. \end{aligned}$$

Using (42) and (43), one can easily show that

$$E \subseteq \bigcup_\alpha E_\alpha^+,$$

from which we obtain

$$\begin{aligned} &\mathbb{P}^N\{w^* \neq 0 \ \wedge \ V(w^*, b^*) \notin I(s^*)\} \\ &\quad = \ \mathbb{Q}^N(E) \\ &\quad \leq \ \mathbb{Q}^N(\cup_\alpha E_\alpha^+) \\ &\quad = \ \lim_{\alpha \to 0} \mathbb{Q}^N(E_\alpha^+) \quad \text{(since } E_\alpha^+ \text{ is increasing as } \alpha \text{ decreases)} \\ &\quad \leq \ \limsup_{\alpha \to 0} \mathbb{Q}^N(E_\alpha) \quad \text{(since } E_\alpha^+ \subseteq E_\alpha) \\ &\quad = \ \limsup_{\alpha \to 0} \mathbb{Q}^N(w^* \neq 0 \ \wedge \ V_\alpha(w_\alpha^*, b_\alpha^*) \notin I(s_\alpha^*)) \\ &\quad \leq \ \limsup_{\alpha \to 0} \mathbb{Q}^N(V_\alpha(w_\alpha^*, b_\alpha^*) \notin I(s_\alpha^*)). \end{aligned}$$

Applying (40) to the last term finally gives

$$\mathbb{P}^N\{w^* \neq 0 \ \wedge \ V(w^*, b^*) \notin I(s^*)\} \leq \beta. \tag{44}$$

To conclude the proof, we have now to account for the realizations of $\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N$ for which $w^* = 0$ and show that

$$\mathbb{P}^N\{w^* = 0 \ \wedge \ V(w^*, b^*) \notin I(s^*)\} \leq 2\beta. \tag{45}$$

In fact, (44) and (45) together give

$$\begin{aligned} &\mathbb{P}^N\{V(w^*, b^*) \notin I(s^*)\} \\ &\quad = \ \mathbb{P}^N\{w^* \neq 0 \ \wedge \ V(w^*, b^*) \notin I(s^*)\} + \ \mathbb{P}^N\{w^* = 0 \ \wedge \ V(w^*, b^*) \notin I(s^*)\} \\ &\quad \leq \ 3\beta, \end{aligned}$$

which is equivalent to the statement of Theorem 5.

To prove (45), first notice that substituting $w^* = 0$ in program (18) gives

$$\begin{aligned} \min_{\substack{b\in\mathbb{R}\\ \xi_i\geq 0, i=1,\ldots,N}} \quad & \rho\sum_{i=1}^{N}\xi_i \\ \text{subject to:} \quad & 1 + y_i b \leq \xi_i, \quad i = 1,\ldots,N, \end{aligned}$$

and a simple direct inspection reveals that at optimum either $b^* = -1$ (when no. of $y_i = 1$ $\geq$ no. of $y_i = -1$; notice that when these two numbers are equal, $b^* = -1$ is enforced by the adopted tie-break rule) or $b^* = 1$ (when no. of $y_i = 1 <$ no. of $y_i = -1$). The analysis is thus split into two sub-cases, namely, $(w^* = 0, b^* = -1)$ and $(w^* = 0, b^* = 1)$, and (45) is obtained by showing that

$$\mathbb{P}^N\{w^* = 0 \ \wedge \ b^* = \odot \ \wedge \ V(w^*, b^*) \notin I(s^*)\} \leq \beta$$

where $\odot$ is either -1 or 1.

The proof for one case is identical to that for the other. Choose thus one, say $(w^* = 0, b^* = -1)$, and consider a version of the heated program (38) where w and b are always (i.e. for all realizations of $\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N$) constrained to take the values 0 and -1, respectively:

$$\begin{aligned} \min_{\substack{w=0, b=-1\\ \xi_i\geq 0, i=1,\ldots,N}} \quad & \|w\|^2 + \rho\sum_{i=1}^{N}\xi_i \\ \text{subject to:} \quad & (1-\alpha z_i) - y_i(\langle w, u_i\rangle - b) \leq \xi_i, \\ & i = 1,\ldots,N, \end{aligned} \tag{46}$$

which is equivalent to

$$\begin{aligned} \min_{\xi_i\geq 0, i=1,\ldots,N} \quad & \rho\sum_{i=1}^{N}\xi_i \\ \text{subject to:} \quad & (1-\alpha z_i) - y_i \leq \xi_i, \quad i = 1,\ldots,N. \end{aligned}$$

Program (46) is quite a peculiar instance of (1), since $x = (w, b)$ belongs to a vector space with null dimensionality. Still, the theory of Section 2 retains its validity. As a matter of fact, (46) has clearly a unique solution, which is

$$\tilde{w}^*_\alpha = 0, \ \tilde{b}^*_\alpha = -1, \ \tilde{\xi}^*_{i,\alpha} = (1-\alpha z_i) - y_i,$$

and it satisfies the non-accumulation Assumption 3 (as shown by (39) with $w = 0$ and $b = -1$). Theorem 1 can therefore be applied to (46) yielding

$$\mathbb{Q}^N\{V_\alpha(\tilde{w}^*_\alpha, \tilde{b}^*_\alpha) \notin I(\tilde{s}^*_\alpha)\} \leq \beta, \tag{47}$$

for all α, where $V_\alpha(\tilde{w}^*_\alpha, \tilde{b}^*_\alpha) = \mathbb{Q}\{(\mathbf{u}, y, z) : (1-\alpha z) - y(\langle \tilde{w}^*_\alpha, \mathbf{u}\rangle - \tilde{b}^*_\alpha) > 0\} = \mathbb{Q}\{(\mathbf{u}, y, z) : y < (1-\alpha z)\}$ and $\tilde{s}^*_\alpha$ is the number of $(\mathbf{u}_i, y_i, z_i)$ for which $(1-\alpha z_i) - y_i(\langle \tilde{w}^*_\alpha, \mathbf{u}_i\rangle - \tilde{b}^*_\alpha) \geq 0$, i.e. for which $y_i \leq (1-\alpha z_i)$.

Recalling that $\alpha = 1/j$, with j any positive integer, that y can be either 1 or -1, and that $z \in [0,1]$, one sees that for all the realizations of $\{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N$ such that $w^* = 0$ and $b^* = -1$ and for all α, it holds that

$$\begin{aligned} V(w^*, b^*) &= V(0,-1) \\ &= \mathbb{P}\{(\mathbf{u}, y) : y < 1\} \\ &= \mathbb{Q}\{(\mathbf{u}, y, z) : y < 1\} \\ &= \mathbb{Q}\{(\mathbf{u}, y, z) : y < (1 - \alpha z)\} \\ &= V_\alpha(\tilde{w}_\alpha^*, \tilde{b}_\alpha^*). \end{aligned}$$

and, with exception of when $z_i = 0$ for some i, which has zero-probability, that

$$\begin{aligned} s^* &= \text{no. of } y_i = -1 \quad \text{(recall how } s^* \text{ is defined when } w^* = 0) \\ &= \text{no. of } y_i \leq (1 - \alpha z_i) \\ &= \tilde{s}_\alpha^*. \end{aligned}$$

Hence, we have

$$\begin{aligned} &\mathbb{P}^N\{w^* = 0 \ \wedge \ b^* = -1 \ \wedge \ V(w^*, b^*) \notin I(s^*)\} \\ &\quad = \mathbb{Q}^N\{w^* = 0 \ \wedge \ b^* = -1 \ \wedge \ V(w^*, b^*) \notin I(s^*)\} \\ &\quad = \mathbb{Q}^N\{w^* = 0 \ \wedge \ b^* = -1 \ \wedge \ V_\alpha(\tilde{w}_\alpha^*, \tilde{b}_\alpha^*) \notin I(\tilde{s}_\alpha^*)\} \\ &\quad \leq \mathbb{Q}^N\{V_\alpha(\tilde{w}_\alpha^*, \tilde{b}_\alpha^*) \notin I(\tilde{s}_\alpha^*)\} \\ &\quad \leq \beta, \end{aligned}$$

which is the sought relation. The same argument applies *mutatis mutandis* for the case $(w^* = 0, b^* = 1)$.

This concludes the proof. ■

References

T. Alamo, R. Tempo, A. Luque, and D. R. Ramirez. Randomized methods for design of uncertain systems: sample complexity and sequential algorithms. *Automatica*, 51: 160–172, 2015.

O. Bousquet, S. Hanneke, S. Moran, and N. Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Proceedings of 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 582–609, Graz, Austria, 2020.

C.J.C. Burges and D.J. Crisp. Uniqueness of the svm solution. In *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, pages 223–229, Denver, CO, 1999.

G.C. Calafiore and M.C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

G.C. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.

M.C. Campi. Classification with guaranteed probability of error. *Machine Learning*, 80: 63–84, 2010.

M.C. Campi and A. Carè. Random convex programs with L_1-regularization: sparsity and generalization. *SIAM Journal on Control and Optimization*, 51(5):3532–3557, 2013.

M.C. Campi and S. Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008. ISSN 1052-6234. doi: http://dx.doi.org/10.1137/07069821X.

M.C. Campi and S. Garatti. Wait-and-judge scenario optimization. *Mathematical Programming*, 167(1):155–189, 2018. doi: https://doi.org/10.1007/s10107-016-1056-9.

M.C. Campi, G. Calafiore, and S. Garatti. Interval predictor models: identification and reliability. *Automatica*, 45(2):382–392, 2009.

A. Carè, S. Garatti, and M.C. Campi. Scenario min-max optimization and the risk of empirical costs. *SIAM Journal on Optimization*, 25(4):2061–2080, 2015.

A. Caré, F.A. Ramponi, and M.C. Campi. A new classification algorithm with guaranteed sensitivity and specificity for medical applications. *IEEE Control Systems Letters*, 2(3): 393–398, 2018.

C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

L.G. Crespo, D.P. Giesy, and S.P. Kenny. Interval predictor models with a formal characterization of uncertainty and reliability. In *Proceedings of the 53rd IEEE Conference on Decision and Control (CDC)*, pages 5991–5996, Los Angeles, CA, USA, 2014.

L.G. Crespo, S.P. Kenny, and D.P. Giesy. Random predictor models for rigorous uncertainty quantification. *International Journal for Uncertainty Quantification*, 5(5):469–489, 2015.

L.G. Crespo, S.P. Kenny, and D.P. Giesy. Interval predictor models with a linear parameter dependency. *Journal of Verification, Validation and Uncertainty Quantification*, 1(2): 1–10, 2016.

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 and Control Letters*, 134:104560, 2019.

A. Falsone, L. Deori, D. Ioli, S. Garatti, and M. Prandini. Optimal disturbance compensation for constrained linear systems operating in stationary conditions: A scenario-based approach. *Automatica*, 110:108537, 2019.

S. Floyd and M. Warmuth. Learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21:269–304, 1995.

S. Garatti and M.C. Campi. Risk and complexity in scenario optimization. *Mathematical Programming*, 2019. doi: https://doi.org/10.1007/s10107-019-01446-4. Published on-line.

S. Garatti, M.C. Campi, and A. Caré. On a class of interval predictor models with universal reliability. *Automatica*, 110:108542, 2019.

T. Graepel, R. Herbrich, and J. Shawe-Taylor. PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59:55–76, 2005.

S. Grammatico, X. Zhang, K. Margellos, P.J. Goulart, and J. Lygeros. A scenario approach for non-convex control design. *IEEE Transactions on Automatic Control*, 61(2):334–345, 2016.

L. Györfi and H. Walk. Nearest neighbor based conformal prediction. *Publications de l'Institut de Statistique de l'Université de Paris*, 63:173–190, 2019. Special issue in honour of Denis Bosq's 80th birthday.

S. Hanneke and A. Kontorovich. Optimality of SVM: novel proofs and tighter bounds. *Theoretical Computer Science*, 796:99–113, 2019.

M.J. Lacerda and L. G. Crespo. Interval predictor models for data with measurement uncertainty. In *Proceedings of the 2017 American Control Conference (ACC)*, pages 1487–1492, Seattle, WA, 2017.

J. Lei, J. Robins, and L. Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108:278–287, 2013.

K. Margellos, P.J. Goulart, and J. Lygeros. On the road between robust optimization and the scenario approach for chance constrained optimization problems. *IEEE Transactions on Automatic Control*, 59(8):2258–2263, 2014.

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*, 60(10):2716–2721, 2015.

M.E. Payton, L.J. Young, and J.H. Young. Bounds for the difference between median and mean of beta and negative binomial distributions. *Metrika*, 36:347–354, 1989.

K. Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. MIT press, 1962.

H. Robbins. A remark on Stirling's formula. *The American Mathematical Monthly*, 62(1), 1955.

G. Schildbach, L. Fagiano, and M. Morari. Randomized solutions to convex programs with multiple chance constraints. *SIAM Journal on Optimization*, 23(4):2479–2501, 2013.

G. Schildbach, L. Fagiano, C. Frei, and M. Morari. The scenario approach for stochastic model predictive control with bounds on closed-loop constraint violations. *Automatica*, 50(12):3009–3018, 2014.

B. Schölkopf and A.J. Smola. *Learning with kernels*. MIT press, 1998.

B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson. Shrinking the tube: A new support vector regression algorithm. In *Advances in Neural Information Processing Systems 11 (NIPS 1998)*, pages 330–336, Denver, CO, 1998.

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

D.M.J. Tax and R.P.W. Duin. Support vector data description. *Machine Learning*, 54: 45–66, 2004.

V. Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92: 349–376, 2013.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, USA, 2005.

X. Wang, F. Chung, and S. Wang. Theoretical analysis for solution of support vector data description. *Neural Networks*, 24:360–369, 2011.

J.S. Welsh and H. Kong. Robust experiment design through randomisation with chance constraints. In *Proceedings of the 18th IFAC World Congress*, Milan, Italy, 2011.

J.S. Welsh and C.R. Rojas. A scenario based approach to robust experiment design. In *Proceedings of the 15th IFAC Symposium on System Identification*, Saint-Malo, France, 2009.

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