

Scenario optimization with relaxation: a new tool for design and application to machine learning problems

Marco C. Campi and Simone Garatti

Abstract—Scenario optimization is by now a well established technique to perform designs in the presence of uncertainty. It relies on domain knowledge integrated with first-hand information that comes from data and generates solutions that are also accompanied by precise statements of reliability. In this paper, following recent developments in [20], we venture beyond the traditional set-up of scenario optimization by analyzing the concept of constraints relaxation. By a solid theoretical underpinning, this new paradigm furnishes fundamental tools to perform designs that meet a proper compromise between robustness and performance. After suitably expanding the scope of constraints relaxation as proposed in [20], we focus on various classical Support Vector methods in machine learning – 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.

I. INTRODUCTION

The scenario approach is a relatively recent, and yet well established, data-driven approach to make reliable designs in the presence of uncertainty. This capability to cope with uncertainty is becoming ever more important in nowadays engineering practice and, after its introduction in the seminal paper [4], the scenario approach has obtained increasing attention as witnessed by many theoretical developments, [9], [1], [10], [30], [13], [23], [38], [14], and it has found application to various fields including control system design, [5], [12], [19], [29], [22], [2], [25], [27], [18], [26], system identification, [37], [7], [36], [17], [21], and machine learning, [6], [8], [24], [15].

Letting x be the vector of design variables (e.g. the parameters of a controller, or those of a regression model or a predictor), the scenario approach builds upon the following two ingredients that are typical of many design problems:

- i. a cost function $c(x)$, which we would like to make as small as possible;
- ii. a family of constraints for x , indexed by δ and expressed as $f(x, \delta) \leq 0$.

The parameter δ that appears in f represents uncertainty and models imprecise knowledge about the environment to which our design will be applied. Each δ corresponds to a potential situation and, for a given δ , constraint $f(x, \delta) \leq 0$ incorporates all restrictions¹ enforced in that situation. Note

M.C. Campi is with the Department of Information Engineering, University of Brescia, via Branze 38, 25123 Brescia, Italy. S. Garatti is with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, piazza L. da Vinci 32, 20133 Milano, Italy. Emails: simone.garatti@polimi.it, marco.campi@unibs.it.

¹When multiple restrictions as expressed by inequalities $f_l(x, \delta) \leq 0$, $l = 1, \dots, L$, are present, one defines $f(x, \delta) = \max_{l=1, \dots, L} f_l(x, \delta)$.

also that $c(x)$ depends on the design variables only. This is with no loss of generality: would the cost depend on δ , one might trace it back to the present setup by introducing a new, equivalent, problem that has an additional variable h where the cost function is $c'(x, h) = h$ and the constraints are $f'(x, h, \delta) \leq 0$ with $f'(x, h, \delta) = \max\{f(x, \delta), c(x, \delta) - h\}$.

One fundamental aspect in the practice of scenario optimization is that one is not required to have at disposal a model for how the uncertainty parameter δ is generated. Indeed δ is modeled as a random element over some probability space $(\Delta, \mathcal{D}, \mathbb{P})$, where Δ and \mathbb{P} remain undefined throughout the algorithmic and theoretical developments of the method. This is practically important since in many applications assuming that Δ and \mathbb{P} are known to the designer is unrealistic: Δ and \mathbb{P} refer to the “real world” and can be truly complex objects in modern engineering 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 make but a few examples). Motivated by this observation, the scenario approach takes data as the primary source of knowledge and, more specifically, it assumes that a sample of instances of δ acquired through experience is available to the designer (these instances are denoted by $\delta_1, \dots, \delta_N$ and called “scenarios”). In this context, the scenario approach maps $\delta_1, \dots, \delta_N$ into a design that tries to minimize $c(x)$ while also satisfying constraints $f(x, \delta) \leq 0$ with high probability.

In practice, cost reduction and constraints satisfaction are often contrasting objectives so that a high level of robustness against constraints violation results in poor performances. A proper trade-off depends on the application at hand and it is important to allow for flexibility in the optimization procedure to accommodate various situations. This was the idea behind the introduction of a new scenario scheme, named scenario optimization with relaxation in Section 5.2 of paper [20], which amounted to consider the following optimization program:

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

The interpretation of (1) is that some scenario constraints $f(x, \delta_i) \leq 0$ can be violated for the purpose of improving the cost 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 can be used to set a suitable trade-off between the original cost and the cost generated by the regret for violating constraints.

Program (1) furnishes a flexible scheme that allows the designer to explore various tentative 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 this process of selection the designer is aided by quantitative tools that describe the quality of the solutions x_ρ^* . Recalling i. and ii., it is natural that the designer is concerned about the achieved cost $c(x_\rho^*)$ and the ensuing risk $V(x_\rho^*)$ where

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

quantifies the probabilistic level of constraints violation. It is important to note that $c(x_\rho^*)$ is readily available to the designer once x_ρ^* has been computed; in contrast, the risk cannot be directly evaluated since its definition involves \mathbb{P} , which is an unknown element of the problem. In [20], it was shown that a tight evaluation of $V(x_\rho^*)$ is possible by adopting the so called wait-&-judge perspective of [11]. Specifically, a certificate on $V(x_\rho^*)$ is obtained from the value taken by an observable quantity s_ρ^* , which is defined as the number of δ_i 's for which $f(x_\rho^*, \delta_i) \geq 0$ (i.e., $s_\rho^* = \text{no. of active constraints} + \text{no. of violated constraints}$). Note also that, the solution x_ρ^* can be reconstructed from these constraints and, hence, s_ρ^* can be interpreted as the *complexity* of the solution.

As is intuitive, the no. of violated constraints alone (empirical risk) is not a valid indicator of the true risk $V(x_\rho^*)$ since optimization generates a bias towards larger risks by drifting the solution against the constraints. The thrust of the result of [20] is that the complexity is instead inescapably linked to $V(x_\rho^*)$ irrespective of the problem at hand, and as such it can be used to always tightly judge the level of risk. It turns out that two scenario solutions with the same empirical risk can have quite different true risks $V(x_\rho^*)$ depending on undisclosed mechanisms by which the satisfaction of some constraints implies the satisfaction of other, unseen, constraints. Nonetheless, it is a universal fact that all these mechanisms are captured by the complexity, which, alone, allows one to derive tight evaluations.

In this paper we build upon the result of [20] and apply it to the important class of Support Vector methods, which have been developed in machine learning for classification and regression problems. Specifically, we consider Support Vector Regression - SVR, [31], [33], Support Vector Machine - SVM, [16], and Support Vector Data Description - SVDD, [34]. It is a fact that all these methods fit quite well the framework of [20] and the dichotomy between cost and constraints satisfaction described above corresponds to the dichotomy between having informative regressors or classifiers and their misprediction or misclassification level for the given data generation mechanism. The main contribution of this paper is to establish all the connections between the general theory of [20] and Support Vector methods, including the necessary adaptations of the theory to the specific setups when required.

It is then shown how the new theory makes a big advance in the reliable usage of Support Vector methods, especially in relation to the long-standing problem of the tuning of hyper-parameters, which is key to obtain good solutions.

Support Vector methods will be dealt with in Section III. For a better understanding of this part, we will first revisit in Section II the theory of [20] and we will present it in a broader setup than that of [20], 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 [20] in the most general setup possible is a second contribution of the present paper. In addition, this section provides a deep asymptotic analysis of the risk when the sample size N tends to infinity. The paper will be closed by a numerical simulation in Section IV, which further clarifies the obtained achievements.

II. RISK ASSESSMENT IN SCENARIO OPTIMIZATION WITH CONSTRAINTS RELAXATION

In this section, we revisit and extend the theory of [20] for the assessment of $V(x_\rho^*)$ (Theorem 1 below). We start by formally stating the assumptions that are required for the derivation. The first specifies the mathematical frame of work, while the second ensures that x_ρ^* 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 $\delta \in \Delta$, $f(x, \delta)$ are convex functionals of x . The scenarios δ_i , $i = 1, \dots, N$, form an independent random sample from $(\Delta, \mathcal{F}, \mathbb{P})$. *

Assumption 2 (existence and uniqueness): Consider optimization problems as in (1) where N is substituted with any index $m = 0, 1, \dots$ and δ_i , $i = 1, \dots, m$, is an independent sample from $(\Delta, \mathcal{F}, \mathbb{P})$. For every m and for every outcome of $(\delta_1, \delta_2, \dots, \delta_m)$, 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 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 $f(x, \delta)$.

²Note 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 , $i = 1, \dots, m$, remain unambiguously determined at optimum by relation $\xi_i^* = f(x^*, \delta_i)$, so that no tie on ξ_i , $i = 1, \dots, m$, can persist after the tie on x is broken.

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

This assumption is connected to the concept of non degeneracy introduced in Definition 3 of [20] and it is often satisfied when δ itself does not accumulate (e.g. when it has density).

We are now ready to present the result that provides a quantitative evaluation of the risk in the context of optimization with constraints relaxation.

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

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

and for $k = N$ consider the polynomial equation

$$1 - \frac{\beta}{6N} \sum_{i=N+1}^{4N} \binom{i}{N} t^{i-N} = 0. \quad (3)$$

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

$$\mathbb{P}^N \{\underline{\epsilon}(s_\rho^*) \leq V(x_\rho^*) \leq \bar{\epsilon}(s_\rho^*)\} \geq 1 - \beta, \quad (4)$$

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

Proof: The proof is easily obtained by noticing that the proof of Theorem 4 in [20], given for the case of optimization over Euclidean spaces, applies *mutatis mutandis* to the present more general setup. *

The main message conveyed by Theorem 1 is that it is possible to construct an interval $[\underline{\epsilon}(s_\rho^*), \bar{\epsilon}(s_\rho^*)]$ where $V(x_\rho^*)$ lies with high confidence $1 - \beta$, and no information on Δ and \mathbb{P} is required in this process (distribution free result). The interval depends on s_ρ^* , which is an observable, and for different values of s_ρ^* we obtain different ranges for $V(x_\rho^*)$, showing that s_ρ^* carries fundamental information for the estimation of $V(x_\rho^*)$. Figure 1 depicts $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$ 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, Section II-A provides asymptotic evaluations and establishes a universal fact that the risk tends to the ratio between complexity and the sample size N as N tends to infinity.

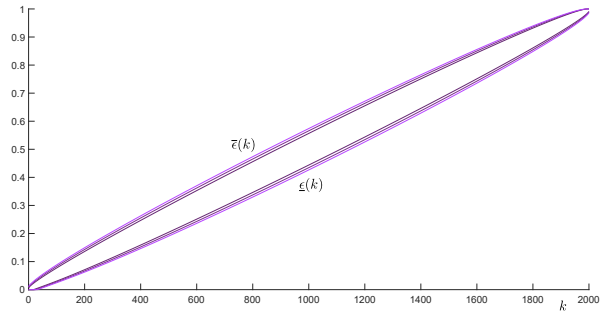


Fig. 1. $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$ for $N = 2000$ and $\beta = 10^{-4}, 10^{-6}, 10^{-8}$.

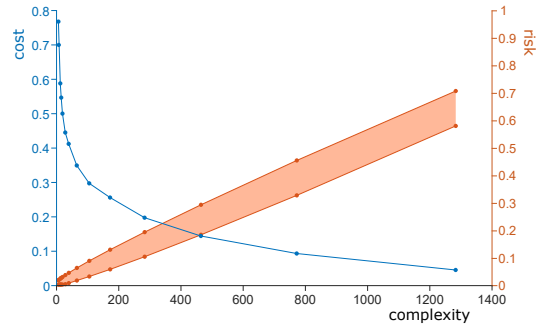


Fig. 2. The cost-risk plot. Dots in the picture correspond to the values of s_ρ^* that have been observed for a range of selections of the parameter ρ .

The typical usage of Theorem 1 is as follows. The designer solves (1) repeatedly for various values of ρ and obtains various solutions x_ρ^* achieving different trade-offs between cost and risk. As ρ varies, the cost is computed, while Theorem 1 allows one to evaluate the risk based on the observed value of the complexity s_ρ^* . In this way, the designer can generate a cost-risk plot like the one depicted in Figure 2, where the cost $c(x_\rho^*)$ and the confidence interval $[\underline{\epsilon}(s_\rho^*), \bar{\epsilon}(s_\rho^*)]$ for $V(x_\rho^*)$ are depicted corresponding to various values of ρ . The user is thus provided with the relevant information to select the solution that achieves the best compromise for the problem at hand.

A. Asymptotic results

We state a theorem that provides explicit bounds for $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$, followed by comments on the asymptotic behavior of these bounds.

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

$$\bar{\epsilon}(k) \leq \frac{k}{N} + C \frac{\sqrt{k} \ln \frac{1}{\beta} + \sqrt{k} \ln k + 1}{N} \quad (5)$$

$$\underline{\epsilon}(k) \geq \frac{k}{N} - C \frac{\sqrt{k} \ln \frac{1}{\beta} + \sqrt{k} \ln k + 1}{N} \quad (6)$$

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 $\bar{\epsilon}(0) \leq (\ln(1/\beta) + 1) \cdot C/N$ and $\underline{\epsilon}(0) \geq 0$. *

Proof: see Appendix A

In (5) and (6), the dependence in β is inversely logarithmic, which shows that “confidence is cheap”. For any fixed k , we see that $\bar{\epsilon}(k)$ 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 .

III. APPLICATION TO SUPPORT VECTOR METHODS

In this section, the general theory for scenario 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.

To ease the notation, we drop from this section onward the subscript ρ in the optimal solution.

A. Support Vector Regression - SVR

Let $\{(\mathbf{u}_i, y_i)\}_{i=1}^N$ 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 extracted independently of the others from a common probability distribution.

Remark 1: 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 only involves the evaluation of inner products in 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 $(\mathbf{m}_1, \dots, \mathbf{m}_n)$ and real values (c_1, \dots, 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. *

In the following, we refer to SVR with adjustable size as described in [31]. For given parameters $\tau, \rho > 0$, consider the optimization program:

$$\begin{aligned} \min_{\substack{w \in \mathcal{U}, \gamma \geq 0, b \in \mathbb{R} \\ \xi_i \geq 0, i=1, \dots, N}} & (\gamma + \tau \|w\|^2) + \rho \sum_{i=1}^N \xi_i \\ \text{subject to:} & |y_i - \langle w, \mathbf{u}_i \rangle - b| - \gamma \leq \xi_i, \quad i = 1, \dots, N. \end{aligned} \quad (7)$$

The cost function in (7) 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 (7), one finds the solution $(w^*, \gamma^*, b^*, \xi_i^*)$, which gives the prediction tube

$$|y - \langle w^*, \mathbf{u} \rangle - b^*| \leq \gamma^*. \quad (8)$$

When a new value $\bar{\mathbf{u}}$ is given, the corresponding output \bar{y} is forecast to be in the tube, that is, in the range of y values such that $|y - \langle w^*, \bar{\mathbf{u}} \rangle - b^*| \leq \gamma^*$. In this process, parameter ρ is used as a tuning knob to adjust the size of the tube against the risk of violating (8), which in this context can be interpreted as the probability of an erroneous prediction (i.e. the actual value \bar{y} is not in the tube). In this context, one should note 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 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 IV for a numerical example.

Remark 2: It is perhaps worth elaborating a bit on what the tube is in case of a lifting in a feature space. As we shall show below, w^* is always given by a linear combination of the points \mathbf{u}_i , say $w^* = \sum_i \alpha^* \mathbf{u}_i$. Substituting in (8) we obtain

$$|y - \sum_i \alpha^* \langle \mathbf{u}_i, \mathbf{u} \rangle - b^*| \leq \gamma^*,$$

which in kernel notation also becomes

$$|y - \sum_i \alpha^* k(\mathbf{m}_i, \mathbf{m}) - b^*| \leq \gamma^*.$$

Throughout, we make the following assumption. *

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

In order to apply the theory from Section II we need to show that the solution to (7) 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 (7) (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 , $i = 1, \dots, m$, does not help satisfy the constraints (note that in the constraints w shows up under the sign of inner product $\langle w, \mathbf{u}_i \rangle$ only), while it increases the cost function (write $w = w_{\mathbf{u}} + w_{\mathbf{u}}^{\perp}$, with $w_{\mathbf{u}} \in \text{span of } \mathbf{u}_i$, $i = 1, \dots, m$, 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, (7) 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 [3] where the same issue is discussed in relation to an algorithmically slightly different, but conceptually identical, problem). Instead, γ^* , b^* and ξ^* 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 $(w^*, \gamma^*, -\bar{b}, \xi_{i,2}^*)$ (keep in mind 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 $\bar{\epsilon}(\cdot)$ as defined in Theorem 1, we have

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

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

B. 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 a lifting into a feature space. See e.g. [34] for a more comprehensive description.

Let $\{\mathbf{p}_i\}_{i=1}^N$ be an independent data set in a Hilbert space \mathcal{P} sampled from a common probability distribution. These points can be raw data or, in complete analogy with the discussion in Remark 1, 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:

$$\begin{aligned} \min_{\substack{c \in \mathcal{P}, \gamma \geq 0 \\ \xi_i \geq 0, i=1, \dots, N}} \quad & \gamma + \rho \sum_{i=1}^N \xi_i \quad (9) \\ \text{subject to:} \quad & \|\mathbf{p}_i - c\|^2 - \gamma \leq \xi_i, \quad i = 1, \dots, N. \end{aligned}$$

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 , $i = 1, \dots, m$, and a solution to (9) 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 [35]; moreover non-uniqueness may only occur when $\rho = 1/M$ for some integer M , refer again to Theorem 3 in [35]. 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 $\forall c, \gamma$:

$$\mathbb{P}\{\|\mathbf{p} - c\|^2 = \gamma\} = 0 \quad \forall c, \gamma. \quad (10)$$

This condition simply requires that probabilistic mass does not accumulate over hyper-spheres and it is formalized in the following assumption.

Assumption 5: Assume that (10) holds. *

We now have the following theorem.

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

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

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

C. 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 the corresponding unseen label to take value -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 [16], [32].

Let $\{(\mathbf{u}_i, y_i)\}_{i=1}^N$ be a data set of independent observations from a common probability distribution, 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 1.

The classifier is obtained by solving the program:

$$\begin{aligned} \min_{\substack{w \in \mathcal{U}, b \in \mathbb{R} \\ \xi_i \geq 0, i=1, \dots, N}} \quad & \|w\|^2 + \rho \sum_{i=1}^N \xi_i \quad (11) \\ \text{subject to:} \quad & 1 - y_i(\langle w, \mathbf{u}_i \rangle - b) \leq \xi_i, \quad i = 1, \dots, N. \end{aligned}$$

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

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

Uniqueness: w^* is unique while b^* may not be, see Theorem 2 in [3]. Break the tie by minimizing $|b + 1|$.³ Similarly to SVR, this returns unique w^* and b^* and the ξ_i^* 's also remain uniquely determined.

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

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

³The 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).

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 (11) returns $w^* = 0$ and $-b^* = 1$ (in case of all labels equal to 1) or $-b^* = -1$ (in case of all labels equal 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 (11); 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 heating and cooling theory is presented in the Appendix, while here below we give 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

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

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

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

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). *

Proof: see Appendix B. *

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:

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

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

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

corresponding to having disagreement between the classifier and the actual sign of y . This condition is more restrictive than constraints violation, and in fact it implies that

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

Importantly, implication is strict and 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

$$\mathbb{P}^N \{ \mathbb{P} \{ (\mathbf{u}, y) : y \text{ is misclassified} \} \leq \bar{\epsilon}(s^*) \} \geq 1 - 3\beta, \quad (12)$$

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. \star

IV. NUMERICAL EXAMPLE

Inspired by the numerical example reported in [31], 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 $y_i = \sin(\pi \mathbf{m}_i) / (\pi \mathbf{m}_i) + e_i$, e_i being extracted from a Laplace distribution with mean $\mu = 0$ and parameter $b = 1$. The kernel trick with Gaussian kernel $\exp(-|\mathbf{m}_k - \mathbf{m}_j|^2)$ was also adopted to expand the input space $[-3, 3]$ into an infinite dimensional feature space in which we cast the optimization in (7). We set $\tau = 0.01$ and then program (7) was repeatedly solved for $\rho = (3/5)^\ell$, $\ell = 0, 1, \dots, 14$. Each time the solution was stored along with the complexity s^* . We set $\beta = 10^{-4}$ and, resorting to Theorem 1 for the calculation of $[\underline{\epsilon}(s^*), \bar{\epsilon}(s^*)]$, we constructed the cost-risk plot, which in fact turned out to be the one in Figure 2. For $\rho = 1$ we obtained the smallest value for s^* , and thereby of the range for the risk, at the expense of a large cost. As ρ increased, s^* also increased monotonically. At the beginning, we had a rapid drop of the cost paired with a moderate increase of the risk. Instead, later, a decrease of cost implied a significant deterioration of the risk. Altogether, this suggested to opt for models corresponding to ρ in the range $(3/5)^7, \dots, (3/5)^{10}$ and, in particular, we chose $\bar{\rho} = (3/5)^9$, yielding $s^* = 105$, corresponding to $[\underline{\epsilon}(s^*), \bar{\epsilon}(s^*)] = [0.032, 0.08]$, and $c(x^*) =$

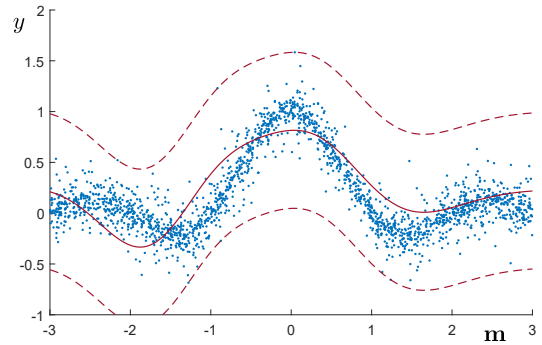


Fig. 3. SVR model for $\rho = 1$.

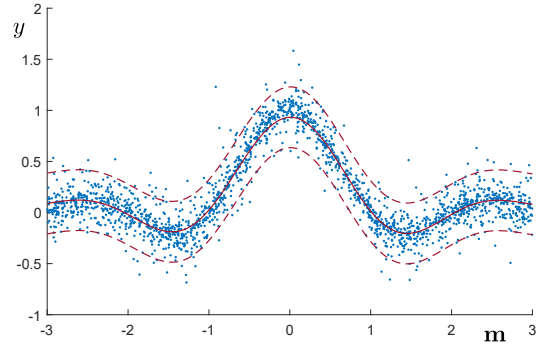


Fig. 4. SVR model for $\rho = (3/5)^9$.

0.30. Since τ was small, γ^* , the size of the tube, was nearly identical to the cost. Figures 3, 4, and 5 depict the models obtained for $\rho = 1$, $\rho = (3/5)^9$, and $\rho = (3/5)^{14}$ and 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 tested the validity of Theorem 1. We kept $\bar{\rho}$ at the value $(3/5)^9$ and solved (7) 200 times, each time drawing a new sample of size 2000. Each solution was tested on 10000 additional random (\mathbf{m}, y) and the risk was evaluated by Monte Carlo techniques. Figure 6 plots the pairs (complexity, risk) obtained in the 200 trials, along with the upper and lower limits given by $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$ when $\beta = 10^{-4}$. Theorem 1 predicts that, on average, the risk is within the prescribed intervals 9999 times out of 10000.

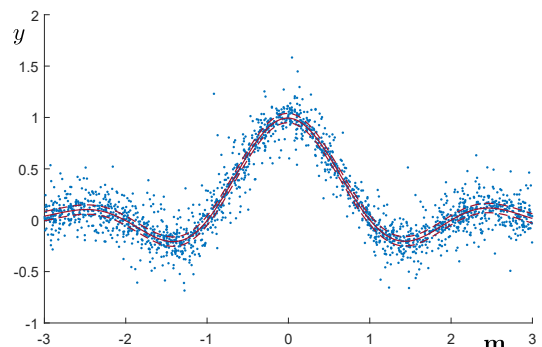


Fig. 5. SVR model for $\rho = (3/5)^{14}$.

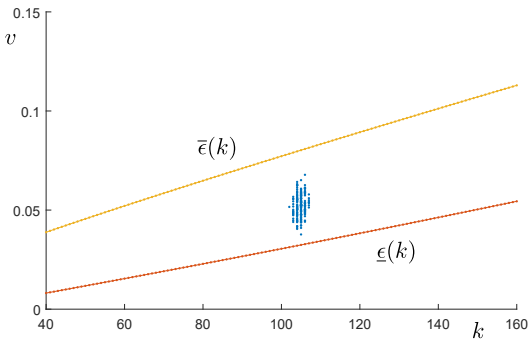


Fig. 6. (complexity,risks) pairs (blue dots) vs. $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$ (continuous dotted lines); $N = 2000$ and $\beta = 10^{-4}$.

As expected, this was the case for all the 200 points in our simulation. A visual inspection also reveals that the spread of the evaluated risks fills well the vertical range given by the theoretical result, a sign that the theoretical result provides tight evaluations in spite of its prerogative of being distribution free.

REFERENCES

- [1] T. Alamo, R. Tempo, and E.F. Camacho. A randomized strategy for probabilistic solutions of uncertain feasibility and optimization problems. *IEEE Transactions on Automatic Control*, 54(11):2545–2559, 2009.
- [2] 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.
- [3] 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.
- [4] G.C. Calafiore and M.C. Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.
- [5] G.C. Calafiore and M.C. Campi. The scenario approach to robust control design. *IEEE Transactions on Automatic Control*, 51(5):742–753, 2006.
- [6] M.C. Campi. Classification with guaranteed probability of error. *Machine Learning*, 80:63–84, 2010.
- [7] M.C. Campi, G. Calafiore, and S. Garatti. Interval predictor models: identification and reliability. *Automatica*, 45(2):382–392, 2009.
- [8] 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.
- [9] 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.
- [10] M.C. Campi and S. Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, 148(2):257–280, 2011.
- [11] M.C. Campi and S. Garatti. Wait-and-judge scenario optimization. *Mathematical Programming*, 167(1):155–189, 2018.
- [12] M.C. Campi, S. Garatti, and M. Prandini. The scenario approach for systems and control design. *Annual Reviews in Control*, 33(2):149 – 157, 2009.
- [13] A. Carè, S. Garatti, and M.C. Campi. FAST - Fast Algorithm for the Scenario Technique. *Operations Research*, 62(3):662–671, 2014.
- [14] 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.
- [15] A. Carè, 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.
- [16] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.
- [17] 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.
- [18] 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, 2019.
- [19] S. Garatti and M.C. Campi. Modulating robustness in control design: principles and algorithms. *IEEE Control Systems Magazine*, 33(2):36–51, 2013.
- [20] S. Garatti and M.C. Campi. Risk and complexity in scenario optimization. *Mathematical Programming*, 2019. Published on-line. DOI: <https://doi.org/10.1007/s10107-019-01446-4>.
- [21] S. Garatti, M.C. Campi, and A. Carè. On a class of interval predictor models with universal reliability. *Automatica*, 110(108542), 2019.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] H.A. Nasir, A. Carè, and E. Weyer. A randomised approach to flood control using value-at-risk. In *Proceedings of the 54th IEEE Conference on Decision and Control (CDC)*, pages 3939–3944, 2015.
- [26] H.A. Nasir, A. Carè, and E. Weyer. A scenario-based stochastic mpc approach for problems with normal and rare operations with an application to rivers. *IEEE Transactions on Control Systems Technology*, 27(4):1397–1410, 2019.
- [27] H.A. Nasir, T. Zhao, A. Carè, Q.J. Wang, and E. Weyer. Efficient river management using stochastic mpc and ensemble forecast of uncertain in-flows. *IFAC-PapersOnLine*, 51(5):37 – 42, 2018. 1st IFAC Workshop on Integrated Assessment Modelling for Environmental Systems IAMES 2018.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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.
- [32] B. Schölkopf and A.J. Smola. *Learning with kernels*.
- [33] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–224, 2004.
- [34] D.M.J. Tax and R.P.W. Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004.
- [35] X. Wang, F. Chung, and S. Wang. Theoretical analysis for solution of support vector data description. *Neural Networks*, 24:360–369, 2011.
- [36] 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.
- [37] 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.
- [38] 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.

A. Proof of Theorem 2

Let $v := 1 - t$. Equation (2) for $k = 0, \dots, N-1$ becomes

$$\begin{aligned} & \frac{\beta}{2N} \sum_{i=k}^{N-1} \binom{i}{k} (1-v)^{i-k} + \frac{\beta}{6N} \sum_{i=N+1}^{4N} \binom{i}{k} (1-v)^{i-k} \\ &= \binom{N}{k} (1-v)^{N-k}. \end{aligned} \quad (13)$$

The fact that (2) has two solutions in $[0, +\infty)$, as stated in Theorem 1, translates into that equation (13) has two solutions in $(-\infty, 1]$, namely $\underline{\epsilon}(k)$ and $\bar{\epsilon}(k)$. Observing that the left-hand side of (13) 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 illustrated in Figure 7. Next consider the following two inequality conditions:

$$\frac{\beta}{2N} \sum_{i=k}^{N-1} \binom{i}{k} (1-v)^{i-k} \geq \binom{N}{k} (1-v)^{N-k}, \quad (14)$$

$$\frac{\beta}{6N} \sum_{i=N+1}^{4N} \binom{i}{k} (1-v)^{i-k} \geq \binom{N}{k} (1-v)^{N-k}. \quad (15)$$

These two inequalities can be used to effectively locate a suitable upper-bound for $\bar{\epsilon}(k)$ (inequality (14)) and lower-bound for $\underline{\epsilon}(k)$ (inequality (15)). This is explained as follows. Take the ratio of the left-hand side over the right-hand side of equation (14):

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

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 $v \rightarrow 1$. Therefore, it picks the value 1 in one and only one point in $(0, 1)$, which shows that equality is attained in (14) for only one value of $v \in (0, 1)$. Hence, the two functions showing up in the left-hand and right-hand sides of (14) are mutually positioned as shown in Figure 7.

Further, it is claimed that any v satisfying (14) is an upper-bound to $\bar{\epsilon}(k)$. Indeed, when moving from equation (13) to (14) we have removed from the left-hand side of (13) a positive term, so shifting to the right the point where equality is achieved in (14); then, owing to the mutual position of the two functions in (14) one immediately sees the correctness of the claim.

The inequality condition (15) can be studied in full analogy to (14) with the only advisory that the role of interval $(0, 1)$ is played by $(1, -\infty)$ when considering the second inequality (15).

Preliminary calculations

To study (14) and (15), we shall use a re-writing of

the left-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

$$\varphi_{H,k}(v) = -\frac{1}{k} \frac{d}{dv} \varphi_{H,k-1}(v). \quad (16)$$

A repeated use (a cumbersome but straightforward exercise) of (16) now gives

$$\varphi_{H,k}(v) = \frac{1 - \sum_{i=0}^k \binom{H}{i} v^i (1-v)^{H-i}}{v^{k+1}} \quad (17)$$

$$= \frac{\sum_{i=k+1}^H \binom{H}{i} v^i (1-v)^{H-i}}{v^{k+1}}. \quad (18)$$

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

Substituting (17) in (14), (14) becomes

$$\frac{\beta}{2} \left(1 - \sum_{i=0}^k \binom{N}{i} v^i (1-v)^{N-i} \right) \geq N \binom{N}{k} v^{k+1} (1-v)^{N-k}. \quad (19)$$

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 [2] and, for any $a > 1$, write:

$$\begin{aligned} & \sum_{i=0}^k \binom{N}{i} v^i (1-v)^{N-i} \\ & \leq a^k \sum_{i=0}^k \binom{N}{i} \left(\frac{v}{a}\right)^i (1-v)^{N-i} \\ & \leq a^k \sum_{i=0}^N \binom{N}{i} \left(\frac{v}{a}\right)^i (1-v)^{N-i} \\ & = a^k \left(1 - v + \frac{v}{a}\right)^N \\ & = (1 - (1-a)) \left(1 - \frac{a-1}{a} v\right)^N \\ & \leq e^{-(1-a)k} e^{-\frac{a-1}{a} v N}, \end{aligned} \quad (20)$$

where the last inequality follows from relation $1 - z \leq e^{-z}$. Similarly,

$$\begin{aligned} & N \binom{N}{k} v^{k+1} (1-v)^{N-k} \\ & \leq (k+1) \binom{N+1}{k+1} v^{k+1} (1-v)^{N+1-(k+1)} \\ & \leq (k+1) \sum_{i=0}^{k+1} \binom{N+1}{i} v^i (1-v)^{N+1-i} \\ & \leq (k+1) e^{-(1-a)(k+1)} e^{-\frac{a-1}{a} v(N+1)} \\ & \leq (k+1) e^{-(1-a)k} e^{-(1-a)k} e^{-\frac{a-1}{a} v N}. \end{aligned} \quad (21)$$

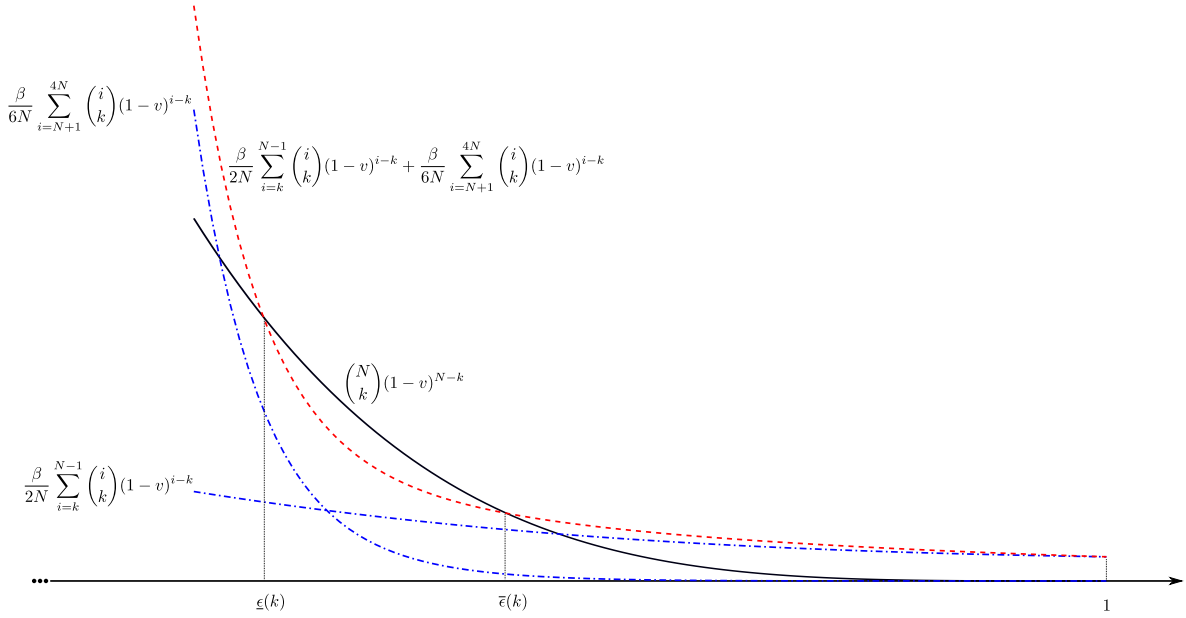


Fig. 7. Graph of functions in (13), (14), and (15).

Suppose now $k > 0$ (the case $k = 0$ will be considered separately) and take $a = 1 + 1/\sqrt{k}$. Using (20) and (21) in (19) yields that any v coming from the inequality

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

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

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

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

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

and the validity of (5) (for $k \neq 0, N$ – recall that we started from equation (2) that holds for $k < N$ and further left behind the case $k=0$) follows by noticing that $\lambda \leq 2$.

Turn now to the case $k = 0, N$.

Case $k = N$ is trivial because $\bar{\epsilon}(N) = 1$, which is clearly in agreement with (5).

As for $k = 0$, go back to (19) and use in it (20) and (21) 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

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

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

Lower bounding $\epsilon(k)$

First, we want to claim that for any k large enough there is a positive v satisfying equation (15). In fact, for $v = 0$ equation (15) 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

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

which is greater than 1 for any

$$k \geq c_1 + c_2 \ln(1/\beta), \quad (22)$$

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 (15).

Using (18) to rewrite the left-hand side of equation (15) as $\sum_{i=N+1}^{4N} \binom{i}{k} (1-v)^{i-k} = \varphi_{4N+1,k}(v) - \varphi_{N+1,k}(v)$, equation

(15) becomes

$$\begin{aligned} & \frac{\beta}{6} \left(\sum_{i=k+1}^{4N+1} \binom{4N+1}{i} v^i (1-v)^{4N+1-i} \right. \\ & \quad \left. - \sum_{i=k+1}^{N+1} \binom{N+1}{i} v^i (1-v)^{N+1-i} \right) \\ & \geq N \binom{N}{k} v^{k+1} (1-v)^{N-k}, \end{aligned} \quad (23)$$

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 $\bar{\epsilon}(k)$, here we can decrease the left-hand side and increase the right-hand side of (23) 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 $v \geq \frac{k+1}{4N+2}$.⁴ Thus, using also the fact $N \binom{N}{k} \leq (k+1) \binom{N+1}{k+1}$, we can take

$$\begin{aligned} & \frac{\beta}{6} \left(\frac{1}{2} - \sum_{i=k+1}^{N+1} \binom{N+1}{i} v^i (1-v)^{N+1-i} \right) \\ & \geq (k+1) \binom{N+1}{k+1} v^{k+1} (1-v)^{N+1-(k+1)} \end{aligned} \quad (24)$$

in place of (23) to obtain a lower bound to $\underline{\epsilon}(k)$ as long as we impose the additional condition that

$$v \geq \frac{k+1}{4N+2}. \quad (25)$$

For any $a > 1$, we now have

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

where the last inequality follows from relation $1+z \leq e^z$. Assume $k > 0$ and take $a = 1 + 1/\sqrt{k}$. Using the above chain of inequalities twice in (24) (for the term in the left-hand side of (24) we use the inequality obtained by comparing the second with the last term in the chain), we obtain the following condition that is more restrictive than (24)

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

⁴This 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, [28].

This inequality is equivalent to

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

which, solved for v , gives

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

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

$$\begin{aligned} v & \leq \frac{k}{N+1} \left(1 - \frac{1}{2\sqrt{k}}\right) \\ & \quad - \frac{\sqrt{k}}{N+1} \left(\ln \frac{12}{\beta} + \ln \left(\frac{\beta}{6} + k + 1\right) \right), \end{aligned} \quad (26)$$

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 (26). Notice also that this equation is valid also for $k = N$ since (3) also leads to (15), 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 (22));
- (ii) $v \geq \frac{k+1}{4N+2}$ (equation (25));
- (iii) $v \leq \frac{k}{N} - g(k, N, \beta)$.

Recall that (iii) makes sense only for $k \neq 0$ (the case $k = 0$ takes care of itself because 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

$$\underline{\epsilon}(k) \geq \frac{k}{N} - C \frac{\sqrt{k} \ln \frac{1}{\beta} + \sqrt{k} \ln k + 1}{N}, \quad (27)$$

with $C = 2C'$. 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 (27) is negative (possibly after enlarging the constant C in (27) to a value that, with a little abuse of notation, we still call C), so that (27) is always a valid lower bound because $\underline{\epsilon}(k)$ is always non-negative. This concludes the proof.

B. Proof of Theorem 5

For analysis purposes, introduce the augmented probability space $(\mathcal{U} \times \mathbb{R}) \times [0, 1]$ 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 $((\mathcal{U} \times \mathbb{R}) \times [0, 1])^N$. Note that this situation traces back to the actual data generation mechanism when $\alpha \rightarrow 0$ because variable z loses its heating role and augmenting $(\mathcal{U} \times \mathbb{R})$ with $[0, 1]$ has no effect.

Suppose we run program (11) with the heated data set, that is, we run

$$\begin{aligned} \min_{\substack{w \in \mathcal{U}, b \in \mathbb{R} \\ \xi_i \geq 0, i=1, \dots, N}} \quad & \|w\|^2 + \rho \sum_{i=1}^N \xi_i \\ \text{subject to:} \quad & (1 - \alpha z_i) - y_i(\langle w, \mathbf{u}_i \rangle - b) \leq \xi_i, \\ & i = 1, \dots, N, \end{aligned} \quad (28)$$

endowed with the same rule adopted in (11) 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:

$$\begin{aligned} & \mathbb{Q}\{(1 - \alpha z) - y(\langle w, \mathbf{u} \rangle - b) = 0\} \\ &= \mathbb{Q}\left\{z = \frac{1 - y(\langle w, \mathbf{u} \rangle - b)}{\alpha}\right\} \\ &= \mathbb{Q}\left\{\mathbb{Q}\left\{z = \frac{1 - y(\langle w, \mathbf{u} \rangle - b)}{\alpha} \mid \mathbf{u}, y\right\}\right\} \\ &= 0. \end{aligned} \quad (29)$$

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

$$\mathbb{Q}^N\{\underline{\epsilon}(s_\alpha^*) \leq V_\alpha(w_\alpha^*, b_\alpha^*) \leq \bar{\epsilon}(s_\alpha^*)\} \geq 1 - \beta, \quad (30)$$

where subscript α indicates that the solution has been obtained from the heated program (28), $V_\alpha(w, b) = \mathbb{Q}\{(\mathbf{u}, y, z) : (1 - \alpha z) - y(\langle w, \mathbf{u} \rangle - b) > 0\}$ 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 (30) 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

$$\begin{aligned} V_0(w^*, b^*) &= (V_0(w^*, b^*) - V_\alpha(w^*, b^*)) \\ &\quad + (V_\alpha(w^*, b^*) - V_\alpha(w_\alpha^*, b_\alpha^*)) \\ &\quad + V_\alpha(w_\alpha^*, b_\alpha^*). \end{aligned} \quad (31)$$

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 $\lim_{\alpha \rightarrow 0}(V_0(w^*, b^*) - V_\alpha(w^*, b^*)) = 0$;

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

Proof of (i): Note that w^* and b^* only depend on the heated training sequence and are treated as deterministic in the calculations that follow to compute 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, $V_0(w^*, b^*) = \mathbb{Q}\{B_0\} = \lim_{\alpha \rightarrow 0} \mathbb{Q}\{B_\alpha\} = \lim_{\alpha \rightarrow 0} V_\alpha(w^*, b^*)$, and claim (i) remains proven.

Proof of (ii): Note that $w_\alpha^* \rightarrow w^*$ and that $b_\alpha^* \rightarrow b^*$ as $\alpha \rightarrow 0$. 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 $\alpha \rightarrow 0$, so that, in the limit, $B_\alpha \Delta B_\alpha^\alpha \subseteq A$ (Δ 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 $(\mathbf{u}, y, z) \notin B_\alpha \Delta B_\alpha^\alpha$ for all $\alpha \leq \bar{\alpha}$. This property in turn implies that $\limsup_{\alpha \rightarrow 0} \mathbb{Q}\{B_\alpha \Delta B_\alpha^\alpha\} \leq \mathbb{Q}\{A\}$ and therefore we have:

$$\begin{aligned} & \limsup_{\alpha \rightarrow 0} |V_\alpha(w^*, b^*) - V_\alpha(w_\alpha^*, b_\alpha^*)| \\ &= \limsup_{\alpha \rightarrow 0} |\mathbb{Q}\{B_\alpha\} - \mathbb{Q}\{B_\alpha^\alpha\}| \\ &\leq \limsup_{\alpha \rightarrow 0} \mathbb{Q}\{B_\alpha \Delta B_\alpha^\alpha\} \\ &\leq \mathbb{Q}\{A\} \\ &= [\text{recall that } w^* \neq 0 \text{ and use Assumption 6}] \\ &= 0. \end{aligned}$$

This completes the proof of (ii).⁵

Using (i) and (ii) in (31), we obtain:

$$\begin{aligned} & \text{for all realizations of } \{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N \text{ such that} \\ & w^* \neq 0, \text{ it holds that} \\ & \lim_{\alpha \rightarrow 0} V_\alpha(w_\alpha^*, b_\alpha^*) = V_0(w^*, b^*). \end{aligned} \quad (32)$$

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

$$\begin{aligned} & \text{with the exception of a zero-probability set,} \\ & \text{for all realizations of } \{(\mathbf{u}_i, y_i, z_i)\}_{i=1}^N \text{ such that} \\ & w^* \neq 0 \text{ it holds that} \\ & \lim_{\alpha \rightarrow 0} s_\alpha^* = s^*. \end{aligned} \quad (33)$$

To see this, note that, when $w^* \neq 0$ and with exception of a zero-probability set, Assumption 6 implies that the

⁵Note 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 $w^* \neq 0$, i.e. when u_i are not picked all equal, we necessarily have $w^* = 1$ and $b^* = 0$, and $V_\alpha(w^*, b^*) = 0$. 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, $\lim_{\alpha \rightarrow 0}(V_\alpha(w^*, b^*) - V_\alpha(w_\alpha^*, b_\alpha^*)) \neq 0$.

(\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 (11), 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 (28) too, which implies that $s_\alpha^* = s^*$ for all α small enough.

Using (30), (32), and (33), we are now ready to establish results that quantify the violation when $w^* \neq 0$.

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

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

Using (32) and (33), one can easily show that

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

from which we obtain

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

Applying (30) to the last term finally gives

$$\mathbb{P}^N \{w^* \neq 0 \wedge V(w^*, b^*) \notin I(s^*)\} \leq \beta. \quad (34)$$

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

$$\mathbb{P}^N \{w^* = 0 \wedge V(w^*, b^*) \notin I(s^*)\} \leq 2\beta. \quad (35)$$

In fact, (34) and (35) together give

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

which is equivalent to the statement of Theorem 5.

To prove (35), first notice that substituting $w^* = 0$ in program (11) gives

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

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 (35) 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 (28) 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:

$$\begin{aligned} &\min_{\substack{w=0, b=-1 \\ \xi_i \geq 0, i=1, \dots, N}} \|w\|^2 + \rho \sum_{i=1}^N \xi_i \\ &\text{subject to: } (1 - \alpha z_i) - y_i(\langle w, \mathbf{u}_i \rangle - b) \leq \xi_i, \\ &\quad i = 1, \dots, N, \end{aligned} \quad (36)$$

which is equivalent to

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

Program (36) is quite a peculiar instance of (1), since $x = (w, b)$ belongs to a vector space with null dimensionality. Still, the theory of Section II retains its validity. As a matter of fact, (36) has clearly a unique solution, which is

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

and it satisfies the non-accumulation Assumption 3 (as shown by (29) with $w = 0$ and $b = -1$). Theorem 1 can therefore be applied to (36) yielding

$$\mathbb{Q}^N \{V_\alpha(\tilde{w}_\alpha^*, \tilde{b}_\alpha^*) \notin I(\tilde{s}_\alpha^*)\} \leq \beta, \quad (37)$$

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}_α^* 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

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

Hence, we have

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

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

This concludes the proof. ★