# Theoretical aspects of robust SVM optimization in Banach spaces and Nash equilibrium interpretation

Mohammed Sbihi, Nicolas Couellan

# Theoretical aspects of robust SVM optimization in Banach spaces and Nash equilibrium interpretation

**Mohammed Sbihi · Nicolas Couellan**

**Abstract** There are many real life applications where data can not be effectively represented in Hilbert spaces and/or where the data points are uncertain. In this context, we address the issue of binary classification in Banach spaces in presence of uncertainty. We show that a number of results from classical support vector machines theory can be appropriately generalized to their robust counterpart in Banach spaces. These include the representer theorem, strong duality for the associated optimization problem as well as their geometrical interpretation. Furthermore, we propose a game theoretical interpretation of the class separation problem when the underlying space is reflexive and smooth. The proposed Nash equilibrium formulation draws connections and emphasizes the interplay between class separation in machine learning and game theory in the general setting of Banach spaces.

**Keywords** Support vector machines · Robust optimization · Nash equilibrium · Duality Mapping

## 1 Introduction

In data science and more specifically in machine learning, the standard assumption is to consider the training data to lie in a Hilbert space. However,

M. Sbihi
ENAC - Université de Toulouse
7 avenue Edouard Belin, 31055, Toulouse, France
E-mail: mohammed.sbihi@enac.fr

N. Couellan
ENAC - Université de Toulouse
7 avenue Edouard Belin, 31055, Toulouse, France

Institut de Mathématiques de Toulouse, Université de Toulouse, UPS IMT, F-31062 Toulouse Cedex 9, France
E-mail: nicolas.couellan@recherche.enac.fr

in some applications where objects are complex such as images, signals, trajectories in robotics or aeronautic, such data representation may turn out to be restrictive or even inefficient. For example, in [1], the authors have shown that Wasserstein-GAN learning achieves better results when considering $L^q$ space embeddings (with $q = 10$) rather than the $L^2$ space. It might be interesting to consider more general representation spaces that better capture and preserve the topological properties of the training samples [14]. More general ambient spaces such as Fréchet spaces have also been considered in [5]. However, as some notion of norm or distance is often required in machine learning methods, Banach spaces are often a better general choice for data embedding. For instance, Banach spaces may be used to model images in a very general manner [1]. Continuous image models that do not rely on the concept of pixel discretization can be regarded as living in the space of measurable functions over the unit square. The use of a specific norm defines a choice of distances between images that can account for specific image features, like the position of edges with Sobolev norms.

In machine learning, Support Vector machines (SVM) [24,31] have been widely used for data classification. Their success is due to sound theoretical foundations and good generalization properties. They address the classification problem by finding the hyperplane that achieves maximum sample margin which leads to minimizing the norm of the classifier parameters. A few studies have demonstrated that classical SVM binary classification formulation may be derived also in non-Euclidean spaces. For example, in [15] a semi-inner-product is considered to formulate a binary classification problem in Banach spaces. In [28], the author also proposed a non-Euclidean setting. General kernels methods in Banach spaces were also investigated in [27,30].

In SVM, it is also common to consider that data is not subject to noise. Data uncertainties are usually not taken into account in classification models, although they occur most of the time. In order to design models that are immune to noise, robust formulations of SVM models have been proposed in the past. Worst case robust optimization formulations [6] have been studied in [13,22], and alternative chance constraint approaches were investigated in [21,26]. In this study, we generalize ideas from the worst case robust method in the context of data lying in a Banach space. The idea is to consider unknown bounded additive noise perturbations of input samples and formulate a robust counterpart training optimization problem when considering worst case scenarios.

Considering data and uncertainties that lie in general Banach spaces, we first propose a theoretical framework that generalizes from an optimization point of view the concept of robust SVM in such spaces. To do so, extending results from robust optimization duality [4], an optimistic dual counterpart problem is derived and robust strong duality is shown to hold under some linearity properties with respect to the uncertainties. The application of these results to the case of robust SVM nicely lead to a representer theorem in Banach spaces. Unlike [15] where a supporting semi-inner product is used in the non Euclidean setting in place of the inner-product, we propose to use

the duality product and show that the representer theorem still holds in this case. Additionally, an uncertain hard margin separation problem and its robust counterpart in Banach spaces are formulated. Furthermore a geometrical perspective of the problem is proposed.

Next, using the geometrical interpretation of the duality analysis proposed before, we formulate the problem of data separation in Banach spaces into the general problem of separating two convex sets representing a class of points. We further show that it can be formulated as a Nash equilibrium problem [16] in a Banach space. Each convex set is seen as a player whose utility is to push the separating plane as far as possible from its points. The proposed interpretation generalizes prior work on supervised classification and Nash-equilibria [12]. This section draws connections and emphasizes the interplay between class separation in machine learning and game theory in the very general setting of Banach spaces.

Our main contributions could be summarized as:

- We propose an extension of the robust SVM formulation when data lie in a Banach space and are subject to unknown bounded uncertainties. Our formulation uses the duality product as opposed to the semi-inner product proposed in prior work.
- We extend results from robust optimization duality and propose an optimistic dual counterpart problem and show robust strong duality under linearity properties with respect to the uncertainties.
- We state and prove a representer theorem for the solution of the robust SVM problem in Banach spaces. This result extends nicely the strong result from classical SVM theory.
- From the geometrical interpretation of the duality results we have derived, we propose a novel formulation of the class separation problem as a Nash-equilibrium problem in Banach spaces. This result emphasizes the relationship between class separation in machine learning and game theory.

The article is organized as follows. Section 2 deals with the robust formulation of the SVM training problem in Banach spaces. Robust optimization duality results are recalled and extended in Section 3. The representer theorem in then stated and proved in Section 4. Section 5 establishes a robust strong duality result and gives its geometrical interpretation. Section 6 describes a game theoretical interpretation of the class separation problem in Banach spaces. Section 7 provides also numerical experiments to illustrate the relationship between the game theoretic formulations and the SVM separation problem. Section 8 terminates with an extension to the non linearly separable case. Section 9 concludes the article.

## 2 Robust SVM optimization in Banach spaces

Let X be a Banach space and $X^*$ be its dual, that is, the space of all real continuous linear functionals on $X$. We recall that $X^*$ is a Banach space endowed

with dual norm defined by

$$\|f\| = \sup_{\|x\|\leq 1} |f(x)|, \quad \forall f \in X^*.$$

There is a natural duality between $X$ and $X^*$ determined by the bilinear functional $\langle \cdot, \cdot \rangle : X \times X^* \to \mathbb{R}$ defined by

$$\langle x, f \rangle = f(x); \forall x \in X, f \in X^*.$$

We first recall the standard SVM methodology [24] to find the maximum margin separating hyperplane between two classes of data points.

Let $x_i \in X$ be a collection of input training vectors for $i = 1, \ldots, m$ and $y_i \in \{-1, 1\}$ be their corresponding labels. If the data are linearly separable, then there exists a linear functional $w \in X^*$ and an offset $b \in \mathbb{R}$ such that $y_i(\langle x_i, w \rangle + b) > 0$ for all $i = 1, \ldots, m$. By rescaling $w$ and $b$, we may assume without loss of generality that the points closest to the hyperplane $H(w, b) := \{x \in X | \langle x, w \rangle + b = 0\}$ satisfy $|\langle x_i, w \rangle + b| = 1$. Thus $H$ may be placed in the canonical form $y_i(\langle x_i, w \rangle + b) \geq 1$, for all $i = 1, \ldots, m$. With this form, the margin of the hyperplane is $\|w\|^{-1}$ (see (21)). We obtain the SVM problem

$$\text{(SVM)} \qquad \begin{array}{ll} \min\limits_{w \in X^*} & \frac{1}{2}\|w\|^2 \\ \text{s.t.} & y_i(\langle x_i, w \rangle + b) \geq 1, \quad i = 1, \ldots, m. \end{array}$$

The classifier is then given by $f(x) = \text{sign}(\langle x, w \rangle + b)$. It is worth mentioning that unlike the formulation given in [15], we use the duality product instead of a semi-inner-product. Considering now instead a set of noisy training vectors $\{\tilde{x}_i \in X, \quad i = 1, \ldots, m\}$ where $\tilde{x}_i = x_i + \delta_i$ for all $i = 1, \ldots, m$ and $\delta_i$ is a random perturbation. This can be captured by the following (noisy SVM) problem

$$\text{(N-SVM)} \qquad \begin{array}{ll} \min\limits_{w \in X^*} & \frac{1}{2}\|w\|^2 \\ \text{s.t.} & y_i(\langle x_i + \delta_i, w \rangle + b) \geq 1, \quad i = 1, \ldots, m. \end{array}$$

Observe that the problem involves the random variable $\delta_i$ and can not be solved as such. Extra knowledge on the perturbations is needed to transform it into a deterministic and numerically solvable problem. In general the perturbation $\delta_i$ is known to reside in some uncertainty set $\Delta_i \subset X$. For instance, [13] considers, when $X = \mathbb{R}^n$, the uncertainty set as $\|\Sigma^{1/2}\delta_i\|_p \leq \gamma_i$, $i = 1, \ldots, m$, where is $\Sigma$ is some positive definite matrix and $p \geq 1$. Various choices of $\Sigma_i$ and $p$ will lead to various types of uncertainties such as for example box-shaped uncertainty ($\|\delta_i\|_\infty \leq \gamma_i$), spherical uncertainty ($\|\delta_i\|_2 \leq \gamma_i$), or ellipsoidal uncertainty ($\delta_i^T \Sigma^{-1} \delta_i \leq \gamma_i^2$). To design a robust model, one has to satisfy the inequality constraint in Problem (N-SVM) for every realizations of $\delta_i$. This can be done by ensuring the constraint in the worst case scenario for $\delta_i$, leading to the following robust counterpart optimization problem:

$$\text{(R-SVM)} \qquad \begin{array}{ll} \min\limits_{w \in X^*} & \frac{1}{2}\|w\|^2 \\ \text{s.t.} & \min\limits_{x_i \in K_i} y_i(\langle x_i, w \rangle + b) \geq 1, \quad i = 1, \ldots, m, \end{array}$$

with $K_i = x_i + \Delta_i$ and with an abuse of notation $\tilde{x}_i$ is denoted by $x_i$.

As previously announced, the aim of this paper is to generalize some known results [8,12,15] from classical SVM to their robust counterpart in Banach spaces. The following section prepares the ground by recalling some facts about robust optimization and by generalizing a robust strong duality result [4] to a Banachic framework tailored to robust SVM related problems.

## 3 A robust optimization detour

We state and adapt in this section some results from robust optimization to our context. Consider a general uncertain optimization problem on some Banach space $\mathcal{B}$

$$\text{(P)} \qquad \inf_{x\in\mathcal{B}} \{f(x) : g_i(x,u_i) \leq 0,\ i = 1,\dots,m\},$$

where $f : \mathcal{B} \to \mathbb{R}$ is a lower-semicontinuous convex function and $g_i : \mathcal{B} \times U_i \to \mathbb{R}, i = 1,\dots,m$, $g_i(\cdot,u_i)$ is convex continuous, $g_i(x,\cdot)$ is upper-semicontinuous and $u_i$ is the uncertain parameter which is only known to reside in certain convex compact uncertainty set $U_i$. Robust optimization, which has emerged as a powerful deterministic approach for studying mathematical programming under uncertainty [4,6,18] associates with the uncertain program (P) its robust counterpart,

$$\text{(R-P)} \qquad \inf_{x\in\mathcal{B}} \left\{ f(x) : \sup_{u_i\in U_i} g_i(x,u_i) \leq 0,\ i = 1,\dots,m \right\},$$

where the uncertain constraints are enforced for every possible value of the parameters within their prescribed uncertainty sets $U_i$. The functions $G_i : \mathcal{B} \ni x \mapsto \sup_{u_i\in U_i} g_i(x,u_i)$, $i = 1,\dots,m$ are convex and continuous as pointwise maxima of convex continuous functions. It is known [3, Theorem 3.9] that under the following Slater condition:

$$\text{There exists a point } x_0 \in \mathcal{B} \text{ such that } G_i(x_0) < 0,\ i = 1,\dots,m, \tag{1}$$

a point $\overline{x} \in \mathcal{B}$ is an optimal solution for (R-P) if and only if there exists $\overline{\lambda} \in \mathbb{R}^m_+$ such that

$$G_i(\overline{x}) \leq 0, \quad i = 1,\dots,m, \tag{2}$$

$$0 \in \partial f(\overline{x}) + \sum_{i=1}^{m} \overline{\lambda}_i \partial G_i(\overline{x}), \tag{3}$$

$$\overline{\lambda}_i G_i(\overline{x}) = 0, \quad i = 1,\dots,m. \tag{4}$$

where for a convex function $h : \mathcal{B} \to \mathbb{R}$, $\partial h(x)$ denotes the Fenchel subdifferential defined by

$$\partial h(x) = \{x^* \in \mathcal{B}^* : h(y) \geq h(x) + \langle y - x, x^*\rangle, \quad \forall y \in \mathcal{B}\}.$$

The following preparatory result refines further the conditions (2)–(4) under linearity assumption with respect to the uncertainties.

**Proposition 1** *Suppose that $U_i$ is a weakly convex subset of some Banach space $\mathcal{C}$ and $g_i(x,\cdot) \in \mathcal{C}^*$ for all $x \in \mathcal{B}$. Under assumption (1), a point $\overline{x} \in \mathcal{B}$ is a minimizer to* (R-P) *if and only if there exist $\overline{\lambda} \in \mathbb{R}^m_+$ and $\overline{u} \in U := \prod_{i=1}^m U_i$ such that*

$$\sup_{u_i \in U_i} g_i(\overline{x}, u_i) \leq 0, \quad i = 1, \ldots, m, \tag{5}$$

$$0 \in \partial f(\overline{x}) + \sum_{i=1}^{m} \overline{\lambda}_i \partial_x g_i(\overline{x}, \overline{u}_i), \tag{6}$$

$$\overline{\lambda}_i g_i(\overline{x}, \overline{u}_i) = 0 \quad i = 1, \ldots, m. \tag{7}$$

*Proof* Let us first show that

$$\partial G_i(x) = \bigcup_{u_i \in U_i(x)} \partial_x g_i(x, u_i) \tag{8}$$

where $U_i(x) = \operatorname{argmax}_{u_i \in U_i} g_i(x, u_i)$. It is known [17] that

$$\partial G_i(x) = \overline{co} \left( \bigcup_{u_i \in U_i(x)} \partial_x g_i(x, u_i) \right)$$

where $\overline{co}$ indicates the closure of the convex hull with respect to weak$*$ topology $\sigma(\mathcal{B}^*, B)$, so proving (8) amounts to prove that $\cup_{x_i \in U_i(x)} \partial_x g_i(x, u_i)$ is convex and weakly* closed. Observe that $U_i(x)$ is convex and a closed subset of $U_i$, hence by the linearity of $g_i$ with respect to $u_i$, it follows that $\bigcup_{u_i \in U_i(x)} \partial_x g_i(x, u_i)$ is convex. So to prove that its weak* closedness, it suffices to prove its sequential weak* closedness. To this end, let $s_n \in \partial_x g_i(x, u_i^n)$, with $u_i^n \in U_i(x)$, converging to some $s \in \mathcal{B}^*$. As $U_i(x)$ is weakly compact, $(s_n)_n$ admits a convergent sub-sequence, still denoted by $(s_n)_n$, converging to some $u_i \in U_i(x)$. By letting $n$ to $+\infty$ in the inequality $g_i(y, u_i^n) \geq g_i(x, u_i^n) + \langle s_n, y - x \rangle$ we get $g_i(y, u_i) \geq g_i(x, u_i) + \langle s, y - x \rangle$, which shows that $s \in \bigcup_{u_i \in U_i(x)} \partial_x g_i(x, u_i)$.

Let us now consider a point $(\overline{x}, \overline{\lambda}, \overline{u})$ satisfying (5)–(7). First note that (2) is not else but (5). For indices $i$ such that $\overline{\lambda}_i = 0$ it is clear that (4) is satisfied. If $\overline{\lambda}_i > 0$, then by (7) $g_i(\overline{x}, \overline{u}_i) = 0$ which combined with (5) yields $0 = g_i(\overline{x}, \overline{u}_i) = \sup_{u_i \in U_i} g_i(\overline{x}, u_i) = G_i(\overline{x})$, so (4) is satisfied. Moreover, by (8) $\partial_x g_i(\overline{x}, \overline{u}_i) \subset G_i(\overline{x})$. Summing over $i$ gives $0 \in \partial f(\overline{x}) + \sum_{i=1}^m \overline{\lambda}_i \partial_x g_i(\overline{x}, \overline{u}_i) \subset \partial f(\overline{x}) + \sum_{i=1}^m \overline{\lambda}_i \partial G_i(\overline{x})$. Consequently, $(\overline{x}, \overline{\lambda})$ satisfy (2)–(4). Let now $(\overline{x}, \overline{\lambda})$ verifying (2)–(4). For such $\overline{\lambda}$, by (3) there exists $v \in \partial f(\overline{x})$, $v_i \in \partial G_i(\overline{x})$ such that $0 = v + \sum_{i=1}^m \lambda_i v_i$. Using (8), for each $i$, there exists $\overline{u}_i \in U_i(\overline{x})$ such that $v_i \in \partial g_i(\overline{x}, \overline{u}_i)$. Finally, we can readily check that the resulting triplet $(\overline{x}, \overline{\lambda}, \overline{u})$ satisfies (5)–(7). □

*Remark 1* Proposition 1 is still valid if we replace the linearity assumption of $g_i(x,\cdot)$ with respect to $u_i$ by (8).

The dual of (R-P) is given by

$$\sup_{\lambda\in\mathbb{R}^m_+}\inf_{x\in\mathcal{B}}\left\{f(x)+\sum_{i=1}^m\lambda_i G_i(x)\right\}$$

which by recalling the definition of $G_i$ becomes

$$\text{(DR-P)}\qquad \sup_{\lambda\in\mathbb{R}^m_+}\inf_{x\in\mathcal{B}}\sup_{u_i\in U_i}\left\{f(x)+\sum_{i=1}^m\lambda_i g_i(x,u_i)\right\}.$$

On the other hand, the uncertain dual of (P) is given by

$$\text{(D-P)}\qquad \sup_{\lambda\in\mathbb{R}^m_+}\inf_{x\in\mathcal{B}}\left\{f(x)+\sum_{i=1}^m\lambda_i g_i(x,u_i)\right\}.$$

The optimistic counterpart of (D-P) is

$$\text{(OD-P)}\qquad \sup_{u\in U,\lambda\in\mathbb{R}^m_+}\inf_{x\in\mathcal{B}}\left\{f(x)+\sum_{i=1}^m\lambda_i g_i(x,u_i)\right\}.$$

By construction, $\inf\text{(R-P)}\geq\sup\text{(DR-P)}\geq\sup\text{(OD-P)}$. The authors in [4] have established, in the case of $\mathcal{B}=\mathbb{R}^n$, that robust strong duality (i.e. $\inf\text{(R-P)}=\max\text{(OD-P)}$) holds between the problems under the Slater condition whenever each $g_i(x,\cdot), i=1,\ldots,m$ is a concave function with respect to $u_i$. In other words, optimizing under the worst case scenario in the primal is the same as optimizing under the best case scenario in the dual ("primal worst equals dual best"). We will establish an analogue result in Banach spaces under some linearity properties with respect to the uncertainties.

By noticing that (R-P) is equivalent to

$$\inf_{x\in\mathcal{B}}\sup_{u\in U,\lambda\in\mathbb{R}^m_+}\left\{f(x)+\sum_{i=1}^m\lambda_i g_i(x,u_i)\right\},$$

(R-P) and (OD-P) can be viewed as dual to each other with $u$ playing the role of an abstract Lagrange multiplier [29, p. 460]. So establishing the strong duality amounts to search the existence of a saddle point of the uncertain lagrangian

$$L:\mathcal{B}\times(\mathbb{R}^m_+\times U)\ni(x;\lambda,u)\mapsto f(x)+\sum_{i=1}^m\lambda_i g_i(x,u_i)$$

with respect to $\mathcal{B}\times(\mathbb{R}^m_+\times U)$, that is a point $(\overline{x};\overline{\lambda},\overline{u})\in\mathcal{B}\times(\mathbb{R}^m_+\times U)$ such that:

$$L(\overline{x};\lambda,u)\leq L(\overline{x};\overline{\lambda},\overline{u})\leq L(x;\overline{\lambda},\overline{u}),\qquad\forall(x;\lambda,u)\in\mathcal{B}\times(\mathbb{R}^m_+\times U).\tag{9}$$

In the sequel we say that $(\overline{x};\overline{\lambda},\overline{u})$ is a solution for the robust primal-optimistic dual (R-P) − (OD-P) pair if $\overline{x}$ is a solution for (R-P) and $(\overline{\lambda},\overline{u})$

is a solution for (OD-P). By [29, Theorem 49.B] $(\overline{x};\overline{\lambda},\overline{u})$ is a solution for the robust primal-optimistic dual pair (R-P) − (OD-P) if and only if $(\overline{x};\overline{\lambda},\overline{u})$ satisfy (9) and in that case the robust strong duality holds, that is $\min$ (R-P) = $\max$ (OD-P).

In the following proposition we link the existence of a saddle point to optimality KKT-like conditions (5)–(7).

**Proposition 2** *Under the assumptions of Proposition 1, a point* $(\overline{x};\overline{\lambda},\overline{u})$ *is a saddle point of* $L$ *with respect to* $\mathcal{B}\times(\mathbb{R}^m_+\times U)$ *if and only if it satisfies (5)–(7).*

*Proof* Suppose that $(\overline{x},\overline{\lambda},\overline{u})$ satisfy (9) then by [29, Theorem 49.B] $\overline{x}$ is a solution (R-P) and $(\overline{\lambda},\overline{u})$ is a solution to (OD-P) so (5) is satisfied. From the right hand of (9), $\overline{x}$ is a minimizer of $L(\cdot,\overline{\lambda},\overline{u})$ and consequently $0\in\partial_x L(\overline{x},\overline{\lambda},\overline{u})=\partial f(\overline{x})+\sum_{i=1}^m\overline{\lambda}_i\partial_x g_i(\overline{x},\overline{u}_i)$, that is (6) is satisfied. It remains to show (7). Consider the no trivial case where $\overline{\lambda}_j>0$. Again from the left-hand side of (9) and by choosing $\lambda$ such that $\lambda_j=\frac{\overline{\lambda}_j}{2}$ and zero otherwise, we get $\frac{\overline{\lambda}_j}{2}g_j(\overline{x},\overline{u}_j)\geq 0$ which combined with $\sup_{u_j\in U}g_j(\overline{x},u_j)\leq 0$ leads to $g_i(\overline{x},\overline{u}_j)=0$. Hence (7) is satisfied. Consider now a point $(\overline{x},\overline{\lambda},\overline{u})$ satisfying (5)–(7). From (6) there exist $d\in\partial f(\overline{x})$, $d_i\in\partial_x g_i(\overline{x},\overline{u}_i)$ for $i=1,\ldots,m$ such that

$$d+\sum_{i=1}^m\overline{\lambda}_i d_i=0.$$

Moreover, we have for all $x\in X$

$$\begin{aligned} f(x)-f(\overline{x})&\geq\langle x-\overline{x},d\rangle,\\ g(x,\overline{u}_i)-g(\overline{x},\overline{u}_i)&\geq\langle x-\overline{x},d_i\rangle,\quad i=1,\ldots,m.\end{aligned}$$

Multiplying by $\overline{\lambda}_i$ $(\geq 0)$ appropriately and summing up all these inequalities,

$$f(x)+\sum_{i=1}^m\overline{\lambda}_i g(x,\overline{u}_i)-(f(\overline{x})+\sum_{i=1}^m\overline{\lambda}_i g(\overline{x},\overline{u}_i))\geq\langle x-\overline{x},d+\sum_{i=1}^m\overline{\lambda}_i d_i\rangle=0$$

that is, $L(\overline{x};\overline{\lambda},\overline{u})\leq L(x;\overline{\lambda},\overline{u})$ for all $x\in\mathcal{B}$. On the other hand, for any $u\in U$ and $\lambda\in\mathbb{R}^m_+$. By (5) then (7)

$$f(\overline{x})+\sum_{i=1}^m\lambda_i g_i(\overline{x},u_j)\leq f(\overline{x})=f(\overline{x})+\sum_{i=1}^m\overline{\lambda}_i g_i(\overline{x},\overline{u}_j)$$

that is, $L(\overline{x};\overline{\lambda},u)\leq L(\overline{x};\overline{\lambda},\overline{u})$ for all $(\lambda,u)\in\mathbb{R}^m_+\times U$. □

## 4 An uncertain representer theorem

In the absence of uncertainty *i.e.* $K_i$ reduced to a single element $x_i$ $i = 1, \ldots, m$, it is shown in [15] that, although posed in an infinite-dimensional space, the optimal solution depends only on the metric relations between the data points. In other words, the problem should not depend on the ambient space in which the data are embedded (such property is known as the Representer Theorem in Hilbert space setting). We shall prove a similar uncertain Representer Theorem in the sense that the optimal solution depends on some realisation of the uncertainty.

We introduce the following notation

- $I_+ = \{i : y_i = +1\}$, $I_- = \{i : y_i = -1\}$ and $I = I_- \cup I_+$,
- $K = \prod_{i=1}^m K_i$,
- $K_\circ = co\left(\bigcup_{i \in I_\circ} K_i\right)$, $\circ \in \{+, -\}$ ($co(A)$ refers to convex hull of $A$)

Note that $K_\circ$ is given by

$$K_\circ = \left\{ \sum_{i \in I_\circ} \alpha_i x_i : x_i \in K_i, 0 \leq \alpha_i \leq 1, i \in I_\circ \text{ and } \sum_{i \in I_\circ} \alpha_i = 1 \right\}$$

and it is weakly compact as soon as each $K_i$ is weakly compact [2, Lemma 5.14]. Moreover we need to introduce the (normalized) duality mapping [10, Definition 4.1] $M : X \to 2^{X^*}$ defined by

$$M(x) = \{x^* \in X^*; \langle x, x^* \rangle = \|x\|^2 = \|x^*\|^2\}. \tag{10}$$

The duality mapping serves as a replacement for the isomorphism $H$ to $H^*$ in the Hilbert case. It is worth noting that $\partial(\frac{1}{2}\|\cdot\|^2)(x) = M(x)$.

We are now able to state the following uncertain Representer Theorem for (R-SVM).

**Theorem 1** *If the uncertainty sets $K_i$, $i = 1, \ldots, m$, are convex and weakly compact and*

$$K_+ \cap K_- = \emptyset, \tag{11}$$

*then* (R-SVM) *admits at least one solution. If $X^*$ is strictly convex then $\overline{w}$ is unique. Moreover, $(\overline{w}, \overline{b})$ is a minimizer to* (R-SVM) *if and only if there exist $\overline{\lambda} \in \mathbb{R}_+^m$ and $\overline{x} \in K$ such that*

$$\max_{x_i \in K} (1 - y_i(\langle x_i, \overline{w} \rangle + \overline{b})) \leq 0 \quad i = 1, \ldots, m, \tag{12}$$

$$\overline{w} \in M(\sum_{i=1}^m y_i \overline{\lambda}_i \overline{x}_i), \tag{13}$$

$$\sum_{i=1}^n y_i \overline{\lambda}_i = 0, \tag{14}$$

$$\overline{\lambda}_i (1 - y_i(\langle \overline{x}_i, \overline{w} \rangle + \overline{b})) = 0 \quad i = 1, \ldots, m. \tag{15}$$

Note that $X^*$ is strictly convex iff for all $w_1, w_2 \in X^*$, $w_1 \neq w_2$, $\|w_1\| = \|w_2\| = 1$, one has $\|\lambda w_1 + (1-\lambda)w_2\| < 1$, $\forall \lambda \in ]0,1[$. In terms of supporting hyperplanes, this property may be expressed as: *distinct boundary points of the closed unit ball have distinct supporting hyperplanes.* This property is equivalent to say that $X$ is smooth [3, Theorem 1.101], that is if for every $x \neq 0$ there exists a unique $x^*$ such that $\|x^*\| = 1$ and $\langle x, x^* \rangle = \|x\|$ or in other words there is exactly one supporting hyperplane through each boundary point of the closed unit ball.

*Proof* Given another Banach space $F$, we will use in the proof the fact that $(X \times F)^*$ is homeomorphic to $X^* \times F^*$ via the linear application $l : X^* \times F^* \ni (h,k) \mapsto h \times k \in (X \times F)^*$, where $(h \times k)(x,y) = h(x) + k(y)$ and $X \times F$ is endowed with the product norm $\|(x,y)\| = \|x\|_X + \|y\|_F$. Let us define the functions $f$, $g_i$ and $G_i$ by $f(w,b) = \frac{1}{2}\|w\|^2$ and $g_i(w,b,x_i) = 1 - y_i(\langle x, w \rangle_{X^*} + b)$ and $G(w,b) = \sup_{x_i \in K_i} g_i(w,b,x_i)$. Let us first show that (11) implies that the Slater condition is satisfied and at the same time the feasible set is not empty. Since $K_-$ and $K_+$ are weakly compact and disjoint, by [19] they can be strictly separated. More precisely there exists $w \in X^*$ such that $\inf_{x \in K_+} \langle x, w \rangle > \sup_{x \in K_-} \langle x, w \rangle$. Let $\alpha, \beta$ such that $\inf_{x \in K_+} \langle x, w \rangle > \alpha > \beta > \sup_{x \in K_-} \langle x, w \rangle$. Set $w_0 = \frac{2w}{\alpha - \beta}$ and $b_0 = -\frac{\alpha + \beta}{\alpha - \beta}$ then $g_i(w_0, b_0) < 0$, $i = 1, \ldots, m$. The objective function $f$ is weakly* lower-semicontinuous and coercive on $X^*$. The feasible set $\cap_{i \in I} \cap_{x_i \in K_i} \{(w,b) \in X^* \times \mathbb{R} : g_i(w,b,x_i) \leq 0\}$ is weakly* closed because $g(\cdot, \cdot, x_i)$ is weakly* continuous. This guarantees the existence of a solution [10, Corollary 1.8.]. The feasible set is convex and the objective function $(\frac{1}{2}\|\cdot\|^2)$ is convex so the set of the minimizers is convex. Given two minimisers $w_1$ and $w_2$, we have $\|w_1\| = \|w_2\| = \|\frac{w_1 + w_2}{2}\|$. When $X^*$ is strictly convex this is possible only if $w_1 = w_2$, which ensures the uniqueness of $w$.

We have $\partial_{w,b} g_i(w,b,x_i) = \{(-y_i x_i, -y_i)\}$ and $\partial f(w,b) = \{(s,0) : s \in M(w)\}$. Applying Proposition 1 and remarking that $\sum_{i=1}^m y_i \overline{\lambda}_i \overline{x}_i \in M(\overline{w})$ is equivalent to $\overline{w} \in M(\sum_{i=1}^m y_i \overline{\lambda}_i \overline{x}_i)$ ends the proof. □

In the $L^p$ case, $1 < p < +\infty$, the duality mapping is single-valued [10, Corollary 4.10] and we obtain the following corollary as in [15].

**Corollary 1** *In the particular case of $X = L^p(\Omega)$, $1 < p < +\infty$, the* (R-SVM) *separating hyperplane admits the expansion* $\overline{w} = \frac{\left|\sum_{i=1}^p \overline{\lambda}_i \overline{x}_i\right|^{p-1} \operatorname{sign}(\sum_{i=1}^p \overline{\lambda}_i \overline{x}_i)}{\|\sum_{i=1}^p \overline{\lambda}_i \overline{x}_i\|^{p-2}} \in L^q(\Omega)$ *(with $\frac{1}{p} + \frac{1}{q} = 1$ and equality in $L^q$ sense).*

## 5 Duality and geometry in Robust SVM classifiers

It is shown in [8] that (SVM) is equivalent to following C-Margin formulation:

$$\text{(CM)} \qquad \begin{array}{ll} \min\limits_{w \in X^*} & \frac{1}{2}\|w\|^2 - (\alpha - \beta) \\ \text{s.t.} & \langle x_i, w \rangle \geq \alpha, \quad i \in I_+, \\ & \langle x_i, w \rangle \leq \beta, \quad i \in I_-. \end{array}$$

whose dual is the problem of finding the closest points in the convex hull of each class. To (CM) we associate the robust version

$$\text{(R-CM)} \qquad \begin{array}{ll} \min\limits_{w\in X^*} & \frac{1}{2}\|w\|^2 - (\alpha-\beta) \\ \text{s.t.} & \min\limits_{x_i\in K_i} \langle x_i, w\rangle \geq \alpha, \quad i\in I_+, \\ & \max\limits_{x_i\in K_i} \langle x_i, w\rangle \leq \beta, \quad i\in I_-. \end{array}$$

Like (R-SVM), the problem (R-CM) admits the following Representer Theorem, the proof of which is identical to Theorem 1 and therefore omitted.

**Theorem 2** *If the uncertainty sets $K_i$, $i=1,\dots,m$, are convex and weakly compact and $K_+\cap K_- = \emptyset$, then* (R-CM) *admits at least one solution. Moreover, $(\overline{w},\overline{\alpha},\overline{\beta})$ is a minimizer to* (R-CM) *if and only if there exists $\overline{\lambda}\in\mathbb{R}^m_+$ and $\overline{x}\in K$ such that*

$$\begin{aligned} &\max_{x_i\in K_i}(\overline{\alpha} - \langle x_i,\overline{w}\rangle) \leq 0, \quad i\in I_+, \\ &\max_{x_i\in K_i}(\langle x_i,\overline{w}\rangle - \overline{\beta}) \leq 0, \quad i\in I_-, \\ &\overline{w}\in M(\sum_{i=1}^m y_i\overline{\lambda}_i\overline{x}_i), \\ &\sum_{i\in I_\circ}\overline{\lambda}_i = 1, \quad \circ\in\{+,-\}, \\ &\overline{\lambda}_i(\overline{\alpha} - \langle\overline{x}_i,\overline{w}\rangle) = 0, \quad i\in I_+, \\ &\overline{\lambda}_i(\overline{\beta} - \langle\overline{x}_i,\overline{w}\rangle) = 0, \quad i\in I_-. \end{aligned}$$

We now extend the duality relationship between (SVM) and (CM) and the resulting geometrical interpretation [8] to their robust counterparts. Define the uncertain Lagrangian $L_1$ on $X^*\times\mathbb{R}\times\mathbb{R}^m_+\times K$ by

$$\begin{aligned} L_1(w,b;\lambda,x) &= \frac{1}{2}\|w\|^2 + \sum_{i=1}^m \lambda_i(1 - y_i(\langle x_i,w\rangle + b)) \\ &= \frac{1}{2}\|w\|^2 - \left\langle\sum_{i=1}^m\lambda_i y_i x_i, w\right\rangle + \sum_{i=1}^m\lambda_i - b\sum_{i=1}^m\lambda_i y_i. \end{aligned}$$

Let us now take a closer look at (OD-SVM)

$$\sup_{(\lambda,x)\in\mathbb{R}^m_+\times K}\ \inf_{(w,b)\in X^*\times\mathbb{R}} L_1(w,b;\lambda,x). \tag{16}$$

Taking the subdifferential of $L_1$ with respect to $w$ and $b$ yields

$$\partial_w L_1(w,b,\lambda,x) = M(w) + \{-\sum_{i=1}^m\lambda_i y_i x_i\}, \tag{17}$$

$$\partial_b L_1(w,b,\lambda,x) = -\sum_{i=1}^m\lambda_i y_i. \tag{18}$$

Letting 0 belong to the subdifferentials (17) and (18) to zero gives

$$\sum_{i=1}^{m} \lambda_i y_i x_i \in M(w), \tag{19}$$

$$\sum_{i=1}^{m} \lambda_i y_i = 0. \tag{20}$$

Substituting (19) and (20) in (16) subject to the relevant constraints yields the dual (OD-SVM) stated as follows

$$\text{(OD-SVM)} \qquad \begin{array}{c} \sup \\ {\scriptstyle (\lambda,x)\in\mathbb{R}^m_+\times K} \end{array} \begin{array}{l} \sum_{i=1}^m \lambda_i - \frac{1}{2}\left\|\sum_{i=1}^m \lambda_i y_i x_i\right\|^2 \\ \text{s.t. } \sum_{i=1}^m \lambda_i y_i = 0. \end{array}$$

By the same arguments and considering $L_2$ on $X^* \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^m_+ \times K$ defined by

$$L_2(w,\alpha,\beta;\lambda,x) = \frac{1}{2}\|w\|^2 - (\alpha-\beta) + \sum_{i\in I_+} \lambda_i(\alpha - \langle x_i, w\rangle) - \sum_{i\in I_-} \lambda_i(\beta - \langle x_i, w\rangle),$$

we get

$$\text{(OD-CM)} \qquad \begin{array}{c} \sup \\ {\scriptstyle (\lambda,x)\in\mathbb{R}^m_+\times K} \end{array} \begin{array}{l} -\frac{1}{2}\left\|\sum_{i=1}^m \lambda_i y_i x_i\right\|^2 \\ \text{s.t. } \sum_{i\in I_\circ} \lambda_i = 1, \quad \circ \in \{+,-\} \end{array}$$

which is not else but the problem of minimizing the (squared) distance between the two convex hulls $K_+$ and $K_-$.

The following theorem states that (R-SVM) and (R-CM) are equivalent.

**Theorem 3** *Assume that the uncertainty sets $K_i$, $i = 1,\dots,m$, are convex and weakly compact and $K_+ \cap K_- = \emptyset$, then*

1. *If $(\overline{w},\overline{b};\overline{\lambda},\overline{x})$ is a solution to the pair* (R-SVM) − (OD-SVM)*, then*

$$\left(\frac{2\overline{w}}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(-1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}; \frac{2\overline{\lambda}}{\sum_{i=1}^m \overline{\lambda}_i}, \overline{x}\right)$$

   *is a solution to the pair* (R-CM) − (OD-CM)*.*
2. *If $(\overline{w},\overline{\alpha},\overline{\beta};\overline{\lambda},\overline{x})$ is a solution to the pair* (R-CM) − (OD-CM) *then*

$$\left(\frac{2\overline{w}}{\overline{\alpha}-\overline{\beta}}, -\frac{\overline{\alpha}+\overline{\beta}}{\overline{\alpha}-\overline{\beta}}; \frac{2\overline{\lambda}}{\overline{\alpha}-\overline{\beta}}, \overline{x}\right)$$

   *is a solution to the pair* (R-SVM) − (OD-SVM)*.*

*Proof* We use Proposition 2 to link the saddles points of $L_1$ and $L_2$. Let us first observe that under the assumptions of the theorem, $\sum_{i=1}^m \overline{\lambda}_i \neq 0$ and $\overline{\alpha} - \overline{\beta} \neq 0$. Suppose that $(\overline{w}, \overline{\alpha}, \overline{\beta}; \overline{\lambda}, \overline{x})$ is a saddle point of $L_2$, then for all $(w, \alpha, \beta; \lambda, x) \in X \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^m_+ \times K$ we have

$$\begin{aligned}
&\frac{1}{2}\|\overline{w}\|^2 - (\overline{\alpha} - \overline{\beta}) + \sum_{i\in I_+} \lambda_i(\overline{\alpha} - \langle x_i, \overline{w}\rangle) - \sum_{i\in I_-} \lambda_i(\overline{\beta} - \langle x_i, \overline{w}\rangle) \\
&\leq \frac{1}{2}\|\overline{w}\|^2 - (\overline{\alpha} - \overline{\beta}) + \sum_{i\in I_+} \overline{\lambda}_i(\overline{\alpha} - \langle \overline{x}_i, \overline{w}\rangle) - \sum_{i\in I_-} \overline{\lambda}_i(\overline{\beta} - \langle \overline{x}_i, \overline{w}\rangle) \\
&\leq \frac{1}{2}\|w\|^2 - (\alpha - \beta) + \sum_{i\in I_+} \overline{\lambda}_i(\alpha - \langle \overline{x}_i, w\rangle) - \sum_{i\in I_-} \overline{\lambda}_i(\beta - \langle \overline{x}_i, w\rangle).
\end{aligned}$$

In particular, by choosing $(\frac{\overline{\alpha}-\overline{\beta}}{2})w$, $\frac{\overline{\alpha}-\overline{\beta}}{2}(1-b)$,$-\frac{\overline{\alpha}-\overline{\beta}}{2}(1+b)$ and $\frac{\overline{\alpha}-\overline{\beta}}{2}\lambda$ instead of $w$, $\alpha$, $\beta$ and $\lambda$ respectively we get that for all $(w, b; \lambda, x) \in X \times \mathbb{R} \times \mathbb{R}^m_+ \times K$

$$\begin{aligned}
&\frac{1}{2}\|\overline{w}\|^2 - (\overline{\alpha} - \overline{\beta}) + \sum_{i\in I_+} \lambda_i \left(\frac{\overline{\alpha} - \overline{\beta}}{2}\right)(\overline{\alpha} - \langle x_i, \overline{w}\rangle) \\
&\qquad\qquad - \sum_{i\in I_-} \lambda_i \left(\frac{\overline{\alpha} - \overline{\beta}}{2}\right)(\overline{\beta} - \langle x_i, \overline{w}\rangle) \\
&\leq \frac{1}{2}\|\overline{w}\|^2 - (\overline{\alpha} - \overline{\beta}) + \sum_{i\in I_+} \overline{\lambda}_i(\overline{\alpha} - \langle \overline{x}_i, \overline{w}\rangle) - \sum_{i\in I_-} \overline{\lambda}_i(\overline{\beta} - \langle \overline{x}_i, \overline{w}\rangle) \\
&\leq \frac{1}{2}\left\|\left(\frac{\overline{\alpha} - \overline{\beta}}{2}\right)w\right\|^2 - \left(\overline{\alpha} - \overline{\beta}\right) \\
&\quad + \sum_{i\in I_+} \overline{\lambda}_i \left(\frac{\overline{\alpha} - \overline{\beta}}{2}\right)(1 - b - \langle \overline{x}_i, w\rangle) + \sum_{i\in I_-} \overline{\lambda}_i \left(\frac{\overline{\alpha} - \overline{\beta}}{2}\right)(1 + b + \langle \overline{x}_i, w\rangle).
\end{aligned}$$

Dividing by $(\frac{\overline{\alpha}-\overline{\beta}}{2})^2$ yields

$$\begin{aligned}
&\frac{1}{2}\left\|\left(\frac{2\overline{w}}{\overline{\alpha} - \overline{\beta}}\right)\right\|^2 + \sum_{i\in I_+} \lambda_i \left(\frac{2\overline{\alpha}}{\overline{\alpha} - \overline{\beta}} - \left\langle x_i, \frac{2\overline{w}}{\overline{\alpha} - \overline{\beta}}\right\rangle\right) \\
&\qquad\qquad - \sum_{i\in I_-} \lambda_i \left(\frac{2\overline{\beta}}{\overline{\alpha} - \overline{\beta}} - \left\langle x_i, \frac{2\overline{w}}{\overline{\alpha} - \overline{\beta}}\right\rangle\right) \\
&\leq \frac{1}{2}\left\|\left(\frac{2\overline{w}}{\overline{\alpha} - \overline{\beta}}\right)\right\|^2 + \sum_{i\in I_+} \frac{2\overline{\lambda}_i}{\overline{\alpha} - \overline{\beta}} \left(\frac{2\overline{\alpha}}{\overline{\alpha} - \overline{\beta}} - \left\langle \overline{x}_i, \frac{2\overline{w}}{\overline{\alpha} - \overline{\beta}}\right\rangle\right) \\
&\qquad\qquad - \sum_{i\in I_-} \frac{2\overline{\lambda}_i}{\overline{\alpha} - \overline{\beta}} \left(\frac{2\overline{\beta}}{\overline{\alpha} - \overline{\beta}} - \left\langle \overline{x}_i, \frac{2\overline{w}}{\overline{\alpha} - \overline{\beta}}\right\rangle\right)
\end{aligned}$$

$$\leq \frac{1}{2}\|w\|^2 + \sum_{i\in I_+} \frac{2\overline{\lambda}_i}{\overline{\alpha}-\overline{\beta}}\left(1-b-\langle \overline{x}_i, w\rangle\right) + \sum_{i\in I_-} \frac{2\overline{\lambda}_i}{\overline{\alpha}-\overline{\beta}}\left(1+b+\langle \overline{x}_i, w\rangle\right).$$

By remarking that $\frac{2\overline{\alpha}}{\overline{\alpha}-\overline{\beta}} = 1 + \frac{\overline{\alpha}+\overline{\beta}}{\overline{\alpha}-\overline{\beta}}$ and $\frac{2\overline{\beta}}{\overline{\alpha}-\overline{\beta}} = -1 + \frac{\overline{\alpha}+\overline{\beta}}{\overline{\alpha}-\overline{\beta}}$ it follows

$$L_1\left(\frac{2\overline{w}}{\overline{\alpha}-\overline{\beta}}, -\frac{\overline{\alpha}+\overline{\beta}}{\overline{\alpha}-\overline{\beta}}; \lambda, x\right) \leq L_1\left(\frac{2\overline{w}}{\overline{\alpha}-\overline{\beta}}, -\frac{\overline{\alpha}+\overline{\beta}}{\overline{\alpha}-\overline{\beta}}; \frac{2\overline{\lambda}}{\overline{\alpha}-\overline{\beta}}, \overline{x}\right) \leq L_1\left(w, b; \frac{2\overline{\lambda}}{\overline{\alpha}-\overline{\beta}}, \overline{x}\right)$$

which means that the point $\left(\frac{2\overline{w}}{\overline{\alpha}-\overline{\beta}}, -\frac{\overline{\alpha}+\overline{\beta}}{\overline{\alpha}-\overline{\beta}}; \frac{2\overline{\lambda}}{\overline{\alpha}-\overline{\beta}}, \overline{x}\right)$ is a saddle point of $L_1$. Conversely, consider a saddle point $(\overline{w}, \overline{b}; \overline{\lambda}, \overline{x})$ of $L_1$, that is for all $(w, b; \lambda, x) \in X \times \mathbb{R} \times \mathbb{R}^m_+ \times K$

$$L_1(\overline{w}, \overline{b}; \lambda, x) \leq L_1(\overline{w}, \overline{b}; \overline{\lambda}, \overline{x}) \leq L_1(w, b; \overline{\lambda}, \overline{x}).$$

Like before by choosing $\frac{\sum_{i=1}^m \overline{\lambda}_i}{2} w$, $\frac{\sum_{i=1}^m \overline{\lambda}_i}{2}(\alpha - \beta)$ and $\frac{\sum_{i=1}^m \overline{\lambda}_i}{2}\lambda$ instead of $w$, $b$ and $\lambda$ and then dividing by $(\frac{\sum_{i=1}^m \overline{\lambda}_i}{2})^2$ we obtain

$$\begin{aligned}
&L_2\left(\frac{2\overline{w}}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(-1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}; \lambda, x\right) \\
&\leq L_2\left(\frac{2\overline{w}}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(-1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}; \frac{2\overline{\lambda}}{\sum_{i=1}^m \overline{\lambda}_i}, \overline{x}\right) \\
&\leq L_2\left(w, \alpha, \beta; \frac{2\overline{\lambda}}{\sum_{i=1}^m \overline{\lambda}_i}, \overline{x}\right)
\end{aligned}$$

which means that the point $\left(\frac{2\overline{w}}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}, \frac{2(-1-\overline{b})}{\sum_{i=1}^m \overline{\lambda}_i}; \frac{2\overline{\lambda}}{\sum_{i=1}^m \overline{\lambda}_i}, \overline{x}\right)$ is a saddle point of $L_2$. □

## 6 Game theoretical interpretation

Based on geometrical properties of class separation in the dual space, a non-cooperative game formulation is given for SVM in [12]. In this section we formulate the problem (OD-CM) as a Nash equilibrium problem for a two-player game. In this game, each player chooses one point from its set and gets a payoff given by the distance between its associated set and an hyperplane defined through the duality mapping that is located at the middle of the segment joining the points chosen by the two players. One may find an interest in such formulation in applications where data privacy is crucial. Indeed, as each player only has knowledge of its own data points, separation can be carried out in a distributed manner where data privacy is preserved.

Given two sets $A, B \subset X$ we denote the distance between $A$ and $B$ by $\text{dist}(A, B) = \inf_{x\in A, y\in B} \|x-y\|$. When $A = \{x\}$, we use the simplified notation

$\mathrm{dist}(x, B)$. The distance from a point to an hyperplane is given by [15, Lemma 1]

$$\mathrm{dist}(x_0, \{x \in X : \langle x, w\rangle - c = 0\}) = \frac{|\langle x_0, w\rangle - c|}{\|w\|}. \tag{21}$$

Moreover, the set of nearest points in $A$ to $x \in X \setminus A$ is denoted $\mathrm{P}_A(x) = \mathrm{argmin}_{y\in A}\|x - y\|$. Suppose that $X$ is smooth which imply that the duality mapping $M$ is single-valued [10, Corollary 4.5].

Consider the following two players game. The player $i$ picks a point $x_i$ in the convex set $C_i$, $i \in \{1, 2\}$. In the context above, $C_1$ (respectively $C_2$) could represent $K_+$ (respectively $K_-$). Then the unique point $w$ of $M(x_1 - x_2)$ is used to define the hyperplane $H(x_1, x_2) := \{x \in X : \langle w, x\rangle = \langle w, \frac{x_1+x_2}{2}\rangle\}$. This hyperplane is halfway between $x_1$ and $x_2$. Indeed

$$\begin{aligned}\mathrm{dist}(x_1, H(x_1, x_2)) &= \frac{\left|\langle w, x_1\rangle - \langle w, \frac{x_1+x_2}{2}\rangle\right|}{\|w\|} \\ &= \frac{|\langle w, x_1 - x_2\rangle|}{2\|w\|} \\ &= \frac{\|x_1 - x_2\|}{2}.\end{aligned}$$

Similarly, $d(x_2, H(x_1, x_2)) = \frac{\|x_1-x_2\|}{2}$. The payoff is defined by

$$v_i(x_1, x_2) := \mathrm{dist}(H(x_1, x_2), C_i), \qquad i \in \{1, 2\}.$$

If $X$ is a Hilbert space then the hyperplane is defined by

$$H(x_1, x_2) = \{x \in X : \langle x_1 - x_2, x - \frac{x_1 + x_2}{2}\rangle = 0\}.$$

This game, denoted by $G$, can be interpreted as if each player was trying to "push" the hyperplane further to himself. The payoff function $v_i$ measures how far the hyperplane is to the player.

A point $(\overline{x}_1, \overline{x}_2)$ is called a Nash Equilibrium (NE) for this game iff

$$\overline{x}_1 \in \mathrm{argmax}_{x_1\in C_1} v_1(x_1, \overline{x}_2) \quad \text{and} \quad \overline{x}_1 \in \mathrm{argmax}_{x_2\in C_2} v_1(\overline{x}_1, x_2).$$

We state the main result of this section.

**Theorem 4** *Let $C_1$ and $C_2$ be two closed convex sets in a reflexive and smooth Banach space $X$. If $C_1 \cap C_2 = \emptyset$, then $(\overline{x}_1, \overline{x}_2)$ is (NE) for $G$ iff $\|\overline{x}_1 - \overline{x}_2\| = \mathrm{dist}(C_1, C_2)$. Moreover, in the that case the payoffs for both players are equal to $\frac{1}{2}\|x_1 - x_2\|$.*

The following lemmas will be used in the proof of Theorem 4.

**Lemma 1** *1. Let $w \neq 0$ be an element from $X^*$ and $H$ the hyperplane $H = \{x : \langle x, w\rangle = c\}$. Then for each pair of points $x_1, x_2 \in X$ strictly separated by $H$, we have*

$$\|x_1 - x_2\| \geq \mathrm{dist}(x_1, H) + \mathrm{dist}(x_2, H).$$

2. *Let $C \subset X$ be a closed convex set and $x_0 \in X$. Suppose that there exists $x_\star \in \mathrm{P}_C(x_0)$ such that $M(x_0 - x_\star) = \{w_\star\}$ then*

$$\langle x - x_\star, w_\star \rangle \leq 0, \qquad \forall x \in C. \tag{22}$$

*Proof* Since $x_1$ and $x_2$ are strictly separated, one of them is located in the positive half-space while the other one is located in the negative half-space. Suppose for example that $\langle x_1, w \rangle < c$ and $\langle x_2, w \rangle > c$. By (21),

$$\begin{aligned} \mathrm{dist}(x_1, H) + \mathrm{dist}(x_2, H) &= \frac{-\langle x_1, w \rangle + c}{\|w\|} + \frac{\langle x_2, w \rangle - c}{\|w\|} \\ &= \frac{-\langle x_2 - x_1, w \rangle}{\|w\|} \\ &\leq \frac{\|x_2 - x_1\| \|w\|}{\|w\|}. \end{aligned}$$

To prove the second item, let $x \in C$ and $\theta \in [0, 1]$ then by the convexity of $C$, $\theta x + (1 - \theta) x_\star \in C$. We have

$$\begin{aligned} 0 &\geq \frac{1}{2}\|x_0 - x_\star\|^2 - \frac{1}{2}\|x_0 - (\theta x + (1 - \theta) x_\star))\|^2 \\ &= \frac{1}{2}\|x_0 - x_\star\|^2 - \frac{1}{2}\|x_0 - x_\star - \theta(x - x_\star)\|^2 \\ &\geq \langle \theta(x - x_\star), w_\theta \rangle \end{aligned}$$

where $w_\theta \in M(x_0 - x_\star - \theta(x - x_\star))$. By dividing by $\theta$ and letting $\theta$ to 0, we obtain the desired inequality since the duality mapping $M$ is norm to weak* upper-semicontinuous [10, Theorem 4.12]. □

**Lemma 2** *Let $C_1$, $C_2$ two closed convex sets of $X$.*

1. *Let $x_1 \in C_1$ and $x_2 \in C_2$ such that $M(x_2 - x_1) = \{w_\star\}$, then*

$$\|x_1 - x_2\| = \mathrm{dist}(C_1, C_2) \Longleftrightarrow x_1 \in \mathrm{P}_{C_1}(x_2) \text{ and } x_2 \in \mathrm{P}_{C_2}(x_1).$$

*Moreover, in that case*

$$\mathrm{dist}(H, C_1) = \mathrm{dist}(H, C_2) = \frac{1}{2}\|x_1 - x_2\|,$$

*where $H$ is the hyperplane defined by $\{x \in X : \langle w_\star, x \rangle = \langle w_\star, \frac{x_1 + x_2}{2} \rangle\}$.*

*Proof* The direct sense is obvious. Consider the reverse one. Since $M(x_1 - x_2)$ is reduced to the single element $-w_\star$, by Lemma 1 we have

$$\langle y_1 - x_1, w_\star \rangle \leq 0, \qquad \forall y_1 \in C_1, \tag{23}$$

$$\langle y_2 - x_2, -w_\star \rangle \leq 0, \qquad \forall y_2 \in C_2. \tag{24}$$

By Lemma 1 again, for all $y_1 \in C_1$, $y_2 \in C_2$ we have

$$\|y_1 - y_2\| \geq \mathrm{dist}(y_1, H) + \mathrm{dist}(y_2, H). \tag{25}$$

Moreover, we have

$$\begin{aligned}\operatorname{dist}(y_1, H) - \frac{1}{2}\|x_1 - x_2\| &= \frac{\langle y_1 - \frac{x_1+x_2}{2}, w_\star\rangle}{\|w_\star\|} - \frac{\langle x_1 - x_2, w_\star\rangle}{2\|w_\star\|} \\ &\geq \frac{\langle y_1 - x_1, w_\star\rangle}{\|w_\star\|} \\ &\geq 0 \qquad \text{(by (23))} \end{aligned} \tag{26}$$

and in the same manner we obtain

$$\operatorname{dist}(y_2, H) - \frac{1}{2}\|x_1 - x_2\| \geq 0. \tag{27}$$

Summing (25),(26) and (27) we get $\|y_1 - y_2\| \geq \|x_1 - x_2\|$ for all $y_1 \in C_1, y_2 \in C_2$ which proves that $\operatorname{dist}(C_1, C_2) = \|x_1 - x_2\|$.
By (26) (respectively (27)), we have $\operatorname{dist}(H, C_1) \geq \frac{1}{2}\|x_1 - x_2\|$ (respectively $\operatorname{dist}(H, C_2) \geq \frac{1}{2}\|x_1 - x_2\|$) and the equality is achieved by $x_1$ (respectively $x_2$) since $\operatorname{dist}(H, x_1) = \frac{1}{2}\|x_1 - x_2\|$ (respectively $\operatorname{dist}(H, x_2) = \frac{1}{2}\|x_1 - x_2\|$). □

*Proof* (of Theorem 4) Suppose that $(\overline{x}_1, \overline{x}_2)$ is a (NE). Let $x_1 \in P_{C_1}(\overline{x}_2)$ and $x_2 \in P_{C_2}(\overline{x}_1)$, their existence is ensured by the reflexivity of $X$ [7]. By Lemma 1 we have

$$\begin{aligned}\|\overline{x}_2 - x_1\| &\geq \operatorname{dist}(\overline{x}_2, H(\overline{x}_1, \overline{x}_2)) + \operatorname{dist}(x_1, H(\overline{x}_1, \overline{x}_2)) \\ &\geq \frac{1}{2}\|\overline{x}_2 - \overline{x}_1\| + \operatorname{dist}(C_1, H(\overline{x}_1, \overline{x}_2)) \\ &\geq \frac{1}{2}\|\overline{x}_2 - \overline{x}_1\| + \operatorname{dist}(C_1, H(x_1, \overline{x}_2)) \qquad (28) \\ &\geq \frac{1}{2}\|\overline{x}_2 - \overline{x}_1\| + \frac{1}{2}\|\overline{x}_2 - x_1\|. \qquad (29)\end{aligned}$$

The inequality (28) comes from the fact that $(\overline{x}_1, \overline{x}_2)$ is a (NE) while (29) from Lemma 1 applied with two convex sets $C_1$ and $\{\overline{x}_2\}$. We obtain from (29) $\|\overline{x}_2 - \overline{x}_1\| \leq \|\overline{x}_2 - x_1\|$ which means that $\overline{x}_1 \in \mathrm{P}_{C_1}(\overline{x}_2)$. Proceeding by the same way we obtain $\overline{x}_2 \in \mathrm{P}_{C_2}(\overline{x}_1)$, that is by Lemma 2 $\|\overline{x}_1 - \overline{x}_2\| = \operatorname{dist}(C_1, C_2)$. Conversely, let$(\overline{x}_1, \overline{x}_2)$ such that $\|\overline{x}_1 - \overline{x}_2\| = \operatorname{dist}(C_1, C_2)$ and suppose by contradiction that $(\overline{x}_1, \overline{x}_2)$ is not (NE). Then there exist $x_2 \in C_2$ (or $x_1 \in C_1$) such that

$$\operatorname{dist}(H(\overline{x}_1, x_2), C_2) > \operatorname{dist}(H(\overline{x}_1, \overline{x}_2), C_2) = \frac{1}{2}\|\overline{x}_1 - \overline{x}_2\|. \tag{30}$$

By Lemma 1 we have

$$\begin{aligned}\|\overline{x}_1 - \overline{x}_2\| &\geq \operatorname{dist}(\overline{x}_1, H(\overline{x}_1, x_2)) + \operatorname{dist}(\overline{x}_2, H(\overline{x}_1, x_2)) \\ &= \frac{1}{2}\|\overline{x}_1 - x_2\| + \operatorname{dist}(\overline{x}_2, H(\overline{x}_1, x_2)) \\ &\geq \frac{1}{2}\|\overline{x}_1 - x_2\| + \operatorname{dist}(C_2, H(\overline{x}_1, x_2)) \\ &> \frac{1}{2}\|\overline{x}_1 - x_2\| + \frac{1}{2}\|\overline{x}_1 - \overline{x}_2\| \qquad \text{(by (30))}.\end{aligned}$$

that is, $\|\overline{x}_1 - \overline{x}_2\| > \|\overline{x}_1 - x_2\|$. This contradicts the fact that $\overline{x}_1$ and $\overline{x}_2$ are the nearest neighbours. □

*Remark 2* Observe that the developments above apply for any convex sets $C_1$ and $C_2$. The sets may not only be convex hulls of data points ($K_+$ and $K_-$) but may arise from other contexts.

## Algorithmic issues

In this part, we propose a numerical algorithm to solve the robust data separation problem based on previous results. The principle is based on the alternating projection method [11,9] to find the minimum distance between the two convex hulls corresponding to the classes. The algorithm can be stated as follows: staring from any $x_1^0 \in C_1$ (or equivalently $x_2^0 \in C_2$), compute the sequences:

$$x_2^n = \mathrm{P}_{C_2}(x_1^n) \text{ and } x_1^{n+1} = \mathrm{P}_{C_1}(x_2^n). \tag{31}$$

Finding the projection $\mathrm{P}_{C_i}(x)$ corresponds to solving a quadratic optimization problem. For the finite dimensional case, when the data uncertainty sets are polytopes, this problem is linearly constrained where as in the case of ellipsoidal uncertainties, the problem is quadratically constrained.

It is known that Algorithm (31) converges and moreover the convergence is finite when $C_1$ and $C_2$ are polytopes [9, Proposition 17].

In the following lemma, we show that computing $\mathrm{P}_{C_i}(x_j^n)$ at each iteration $n$ corresponds to finding the best response for player $i$ when the strategy of player $j$ is fixed to $x_j^n$.

**Lemma 3** *If player $j$ chooses the strategy $\bar{x}_j$, the best response strategy for the other player $i$ is given by the projection of $\bar{x}_j$ onto $C_i$.*

*Proof* Without loss of generality assume that $i = 2$. Suppose for the sake of contradiction that $\mathrm{P}_{C_2}(\bar{x}_1)$, noted $\bar{x}_2$, is not a best response strategy for player 2. Then there exists $x_2 \in C_2$ such that

$$\mathrm{dist}(H(\overline{x}_1, x_2), C_2) > \mathrm{dist}(H(\overline{x}_1, \overline{x}_2), C_2) = \frac{1}{2}\|\bar{x}_1 - \bar{x}_2\|. \tag{32}$$

Since $\bar{x}_1$ and $\bar{x}_2$ are separated by the hyperplane $H(\overline{x}_1, x_2)$, by Lemma 1 we have

$$\begin{aligned}
\|\bar{x}_1 - \bar{x}_2\| &\geq \mathrm{dist}(\bar{x}_1, H(\bar{x}_1, x_2)) + \mathrm{dist}(\bar{x}_2, H(\bar{x}_1, x_2)) \\
&= \frac{1}{2}\|\bar{x}_1 - x_2\| + \mathrm{dist}(\bar{x}_2, H(\bar{x}_1, x_2)) \\
&\geq \frac{1}{2}\|\bar{x}_1 - x_2\| + \mathrm{dist}(C_2, H(\bar{x}_1, x_2)) \\
&> \frac{1}{2}\|\bar{x}_1 - x_2\| + \frac{1}{2}\|\bar{x}_1 - \bar{x}_2\| \qquad \text{(by (32))}.
\end{aligned}$$

So, $\|\overline{x}_1 - \overline{x}_2\| > \|\overline{x}_1 - x_2\|$ which contradicts the fact that $\bar{x}_2 \in \mathrm{P}_{C_2}(\bar{x}_1)$. □

The lemma 3 combined with Theorem 4 ensures that Algorithm (31) converges towards the Nash equilibrium of the game $G$.

## 7 Numerical experiments

In this section, we provide some experiments to illustrate numerically the relationship between the search for a Nash equilibrium in the game defined above and the optimal separation using SVM.

The experiments are conducted on 2D samples in order to facilitate the visualization of data. For simplicity, we also considered that the data samples are not subject to uncertainty. This can be done without loss of generality in the illustrations that are provided. Indeed, in the following, when uncertainties are considered, the convex hulls are enlarged by the uncertainty radius at the boundary of the convex hull. New convex hull simplices have to be considered but insights and illustrations discussed below are similar.
In the experiments, $m$ takes the values of 100, 1000 and 10000. Using `Python` [23], for each value of $m$, the convex hull of the two clusters $C_1$ and $C_2$ are computed using the `ConvexHull` method of the `scipy.spatial` [25] package. The optimal separation given by the Nash equilibrium in Theorem 4 is then computed and shown on the second row of Figure 1. In parallel, the optimal separation found by the SVM method is found using the `SVC` method of the `scikit-learn` [20] package. The optimal separation hyperplane is shown on the third row of Figure 1. One can clearly see that the hyperplane computed through the game theoretic method and the SVM method are identical. This confirms and illustrates the duality between the game theoretic and the SVM formulation of the data separation problem. Figure 1 also provides in the first row, an intermediate situation in which the utiliy of each player (distance to the hyperplane) is computed but not optimal. The Nash equilibrium is not reached and it can be seen clearly that the optimal separating hyperplane has not been found.

## 8 Extension to the non separable case

Let us now suppose that $K_-$ and $K_+$ are non-linearly separable. A linear robust soft margin SVM training can be formulated by using slack variables which measure the degree of misclassification of the observations leading to the following relaxed version

$$\text{(R-SVM}(C)\text{)} \qquad \begin{array}{ll} \min\limits_{(w,b,\xi)\in X^*\times\mathbb{R}\times\mathbb{R}^m_+} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i \\ \text{s.t.} & \min\limits_{x_i\in K_i} y_i(\langle x_i, w\rangle + b) \geq 1-\xi_i, \quad i=1,\dots,m, \end{array}$$

where $C > 0$ is a problem specific constant controlling the trade-off between margin (generalisation) and classification. The optimistic counterpart of its

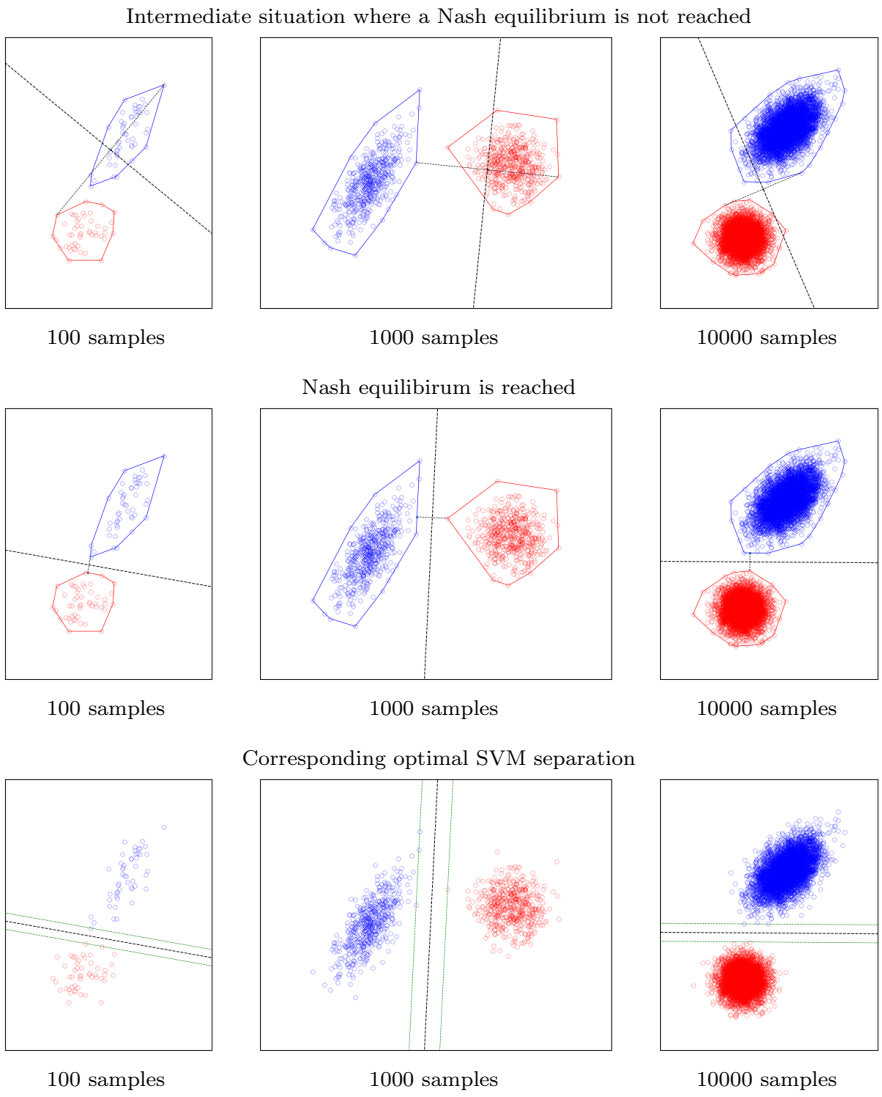


**Fig. 1** Separation of clusters of randomly generated 2D samples

corresponding uncertain dual is

$$\text{(OD-SVM}(C)\text{)} \qquad \begin{array}{ll} \sup\limits_{(\lambda,x)\in\mathbb{R}^m_+\times K} & \sum_{i=1}^m \lambda_i - \frac{1}{2}\left\|\sum_{i=1}^m \lambda_i y_i x_i\right\|^2 \\ \text{s.t.} & \sum_{i=1}^m \lambda_i y_i = 0, \\ & \lambda_i \leq C,\ i = 1,\ldots,m. \end{array}$$

In a similar way we formulate the relaxed version of (R-CM)

$$\text{(R-CM}(D)\text{)} \qquad \begin{array}{ll} \min\limits_{(w,\alpha,\beta,\xi)\in X^*\times\mathbb{R}\times\mathbb{R}\times\mathbb{R}^m_+} & \frac{1}{2}\|w\|^2-(\alpha-\beta)+D\sum_{i=1}^m \xi_i \\ \text{s.t.} & \min\limits_{x_i\in K_i}\langle x_i,w\rangle \geq \alpha-\xi_i, \quad i\in I_+, \\ & \max\limits_{x_i\in K_i}\langle x_i,w\rangle \leq \beta+\xi_i, \quad i\in I_- \end{array}$$

whose optimistic counterpart of its corresponding uncertain dual is

$$\text{(OD-CM}(D)\text{)} \qquad \begin{array}{ll} \sup\limits_{(\lambda,x)\in\mathbb{R}^m_+\times K} & -\frac{1}{2}\left\|\sum_{i=1}^m \lambda_i y_i x_i\right\|^2 \\ \text{s.t.} & \sum_{i\in I_\circ}\lambda_i = 1, \quad \circ\in\{+,-\}, \\ & \lambda_i \leq D,\ i=1,\ldots,m. \end{array}$$

This is in fact not else but the problem of minimizing the (squared) distance between the two convex sets

$$K_\circ(D)=\left\{\sum_{i\in I_\circ}\alpha_i x_i : x_i\in K_i, 0\leq\alpha_i\leq D, i\in I_\circ \text{ and } \sum_{i\in I_\circ}\alpha_i=1\right\}, \circ\in\{-,+\},$$

corresponding to Reduced Convex Hull following the terminology of [8]. Under this form, it is clear that reducing $D$ sufficiently will ensure separability of the problem. The results established in the separable case can, by almost similar arguments, be extended to this non-separable case. We can show that optimizing R-SVM($C$) is equivalent to optimizing R-CM($D$). The parameters $C$ and $D$ are related by multiplication of a constant factor as shown by the following theorem.

**Theorem 5** *Assume that the uncertainty sets $K_i$, $i=1,\ldots,m$, are convex and weakly compact, then*

1. *If $(\overline{w},\overline{b},\overline{\xi};\overline{\lambda},\overline{x})$ is solution for* (R-SVM($C$)) − (OD-SVM($C$)), *with $\overline{w}\neq 0$, then*
$$\left(\frac{2\overline{w}}{\sum_{i=1}^m\overline{\lambda}_i},\frac{2(1-\overline{b})}{\sum_{i=1}^m\overline{\lambda}_i},\frac{2(-1-\overline{b})}{\sum_{i=1}^m\overline{\lambda}_i},\frac{2\overline{\xi}}{\sum_{i=1}^m\overline{\lambda}_i};\frac{2\overline{\lambda}}{\sum_{i=1}^m\overline{\lambda}_i},\overline{x}\right)$$
*is a solution to* (R-CM($\frac{2C}{\sum_{i=1}^m\overline{\lambda}_i}$)) − (OD-CM($\frac{2C}{\sum_{i=1}^m\overline{\lambda}_i}$)).
2. *If $(\overline{w},\overline{\alpha},\overline{\beta},\overline{\xi};\overline{\lambda},\overline{x})$ is solution to* (R-CM($D$)) − (OD-CM($D$)), *with $\overline{w}\neq 0$, then*
$$\left(\frac{2\overline{w}}{\overline{\alpha}-\overline{\beta}},-\frac{\overline{\alpha}+\overline{\beta}}{\overline{\alpha}-\overline{\beta}},\frac{2\overline{\xi}}{\overline{\alpha}-\overline{\beta}};\frac{2\overline{\lambda}}{\overline{\alpha}-\overline{\beta}},\overline{x}\right)$$
*is a solution for* (R-SVM($\frac{2D}{\overline{\alpha}-\overline{\beta}}$)) − (OD-SVM($\frac{2D}{\overline{\alpha}-\overline{\beta}}$)).

This theorem has been established in [8] in the finite dimensional case when data are not subject to uncertainties. The authors also provide geometrical insights of the theorem. The interpretation remains valid for the uncertain case and also the infinite setting.

## 9 Conclusion

This theoretical analysis is an additional step towards the generalization of formulations of binary classification problems in Banach spaces. In [15], it had already been shown that classical SVM formulations nicely extends to Banach spaces by the use of semi-inner products. The authors had shown that most of hard margin separation results in Hilbert spaces remain valid in the non Euclidean setting when considering an appropriate alternative to inner products. In our study, we show that using the duality product, we not only also retrieve the binary classification formulation but robust formulations can also be derived when data uncertainties lie in Banach spaces. Furthermore, using the classification formulation based on the duality product, we show that game theoretic interpretations can also be made. This bridge between game theory and classification of complex data (represented in Banach spaces rather than Hilbert spaces) opens new opportunities for exploiting theoretical and numerical results from both worlds.

**Acknowledgements** The authors would like to thank Prof. Jean-Baptiste Hiriart-Urruty (Université Toulouse III- Paul Sabatier, France) for his insightful and constructive discussions about this research.

## Funding

This work has partially benefited from the AI Interdisciplinary Institute ANITI. ANITI is funded by the French "Investing for the Future - PIA3" program under the Grant agreement # ANR-19-PI3A-0004.

## References

1. Adler, J., Lunz, S.: Banach Wasserstein GAN, In Advances in Neural Information Processing Systems 32 (NIPS 2018), 2018.
2. Aliprantis, C., Border, K.C.: Infinite-dimensional analysis, A hitchhiker's guide. Springer-Verlag, Berlin (1999). https://doi.org/10.1007/978-3-662-03961-8
3. Barbu, V., Precupanu, T.: Convexity and optimization in Banach spaces, Springer, Dordrecht (2012).https://doi.org/10.1007/978-94-007-2247-7
4. Beck, A., Ben-Tal, A.: Duality in robust optimization: primal worst equals dual best. Oper. Res. Lett. 37(1) 1–6 (2009). https://doi.org/10.1016/j.orl.2008.09.010
5. Benth, F.E., Detering, N., Galimberti, L. : Neural networks in Fréchet spaces. Ann Math Artif Intell 91, 75–103 (2023). https://doi.org/10.1007/s10472-022-09824-z
6. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.S.: Robust Optimization, Princeton Series in Applied Mathematics, Princeton University Press (2009). https://doi.org/10.1515/9781400831050
7. Borwein, J.M., Fitzpatrick, S.: Existence of nearest points in Banach spaces. Canad. J. Math. 41(4), 702–720 (1989). https://doi.org/10.4153/CJM-1989-032-7
8. Bredensteiner, E.J., Bennett, K.P.: Duality and geometry in SVM classifiers. Proceedings of the 17th International Conference on Machine Learning. 57–64 (2000)
9. Bui, H.T., Loxton, R., Moeini, A.: A note on the finite convergence of alternating projections, Operations Research Letters, 49 (3),431-438 (2021)

10. Cioranescu, I.: Geometry of Banach spaces, duality mappings and nonlinear problems, Kluwer Academic Publishers Group, Dordrecht (1990). https://doi.org/10.1007/978-94-009-2121-4
11. Cheney, W., Goldstein, A. A.: Proximity Maps for Convex Sets. Proceedings of the American Mathematical Society, 10(3), 448–450 (1959)
12. Couellan, N.: A note on supervised classification and Nash-equilibrium problems. RAIRO Oper. Res. 51(2), 329–341 (2017). https://doi.org/10.1051/ro/2016024
13. Couellan, N., Jan, S.: Feature uncertainty bounds for explicit feature maps and large robust nonlinear SVM classifiers. Ann. Math. Artif. Intell. 88(1-3), 269–289 (2020). https://doi.org/10.1007/s10472-019-09676-0
14. Delahaye, D.,Puechmorel, S., Alam, S., Féron, E.:Trajectory Mathematical Distance Applied to Airspace Major Flows Extraction. EIWAC 2017, 5th ENRI International Workshop on ATM/CNS, Tokyo, Japan, Lecture Notes in Electrical Engineering, Spinger, 555, 51–67 (2017).
15. Der, R., Lee, D.: Large-margin classification in Banach space. Journal of Machine Learning Research - Proceedings Track. 2, 91–98
16. Drew, F., Tirole, J.: Game theory. MIT press, 1991.
17. Hantoute, A., López, M. A., Zălinescu, C.: Subdifferential calculus rules in convex analysis: a unifying approach via pointwise supremum functions. SIAM J. Optim. 19(2), 863–882 2008. https://doi.org/10.1137/070700413
18. Jeyakumar, V., Li, G.Y.: Strong duality in robust convex programming: complete characterizations. SIAM J. Optim. 20(6), 3384–3407 (2010). https://doi.org/10.1137/100791841
19. Klee, J.V.L.: Convex sets in linear spaces. Duke Math. J. 18, 443–466 (1951). http://projecteuclid.org/euclid.dmj/1077476574
20. Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830 (2011)
21. Peng, S., Canessa, G., Allen-Zhao, Z.: Chance constrained conic-segmentation support vector machine with uncertain data. Ann Math Artif Intell (2023). https://doi.org/10.1007/s10472-022-09822-1
22. Trafalis, T.B., Gilbert, R.C.: Robust classification and regression using support vector machines. European J. Oper. Res. 173(3), 893–909 (2006). https://doi.org/10.1016/j.ejor.2005.07.024
23. Van Rossum, G, rake Jr., , Fred L.: Python reference manual.,Centrum voor Wiskunde en Informatica Amsterdam (1995)
24. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag, New York (2000). https://doi.org/10.1007/978-1-4757-3264-1
25. Virtanen, P., Gommers, R., Oliphant, T., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van der Walt, S., Brett, M., Wilson, J., Millman, K., Mayorov, N., Andrew R. J.,Jones, E., Kern, R. Larson, E., Carey, C J.,Polat, I., Feng, Y., Moore, E., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C., Archibald, A., Ribeiro, A., Pedregosa, F.van Mulbregt, P.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,Nature Methods, 17,261–272, (2020)
26. Wang, X., Fan, N., Pardalos, P.M.: Robust chance-constrained support vector machines with second-order moment information. Ann Oper Res 263, 45–68 (2018). https://doi.org/10.1007/s10479-015-2039-6
27. Xu, Y., Ye, Q.: Generalized Mercer kernels and reproducing kernel Banach spaces. Mem. Amer. Math. Soc. 258(1243), vi+122, (2019). https://doi.org/10.1090/memo/1243
28. Ying, L., Qi,Y.: Support vector machine classifiers by non-euclidean margins. Mathematical Foundations of Computing. 3(4), 279–300 (2020).
29. Zeidler, E.: Nonlinear functional analysis and its applications. III, Variational methods and optimization, Translated from the German by Leo F. Boron, Springer-Verlag, New York (1985). https://doi.org/10.1007/978-1-4612-5020-3
30. Zhang, H., Xu, Y., Zhang, J.: Reproducing kernel Banach spaces for machine learning. J. Mach. Learn. Res. 10, 2741–2775 (2009). https://doi.org/10.1109/IJCNN.2009.5179093
31. Schölkopf, B., Smola, A.: Learning with Kernels. MIT, Cambridge (2002).