Large-Margin Classification in Banach Spaces

Ricky Der
Department of Mathematics
University of Pennsylvania
Philadelphia, PA 19103

Daniel Lee
Department of Electrical Engineering
University of Pennsylvania
Philadelphia, PA 19103

Abstract

We propose a framework for dealing with binary hard-margin classification in Banach spaces, centering on the use of a supporting semi-inner-product (s.i.p.) taking the place of an inner-product in Hilbert spaces. The theory of semi-inner-product spaces allows for a geometric, Hilbert-like formulation of the problems, and we show that a surprising number of results from the Euclidean case can be appropriately generalised. These include the Representer theorem, convexity of the associated optimization programs, and even, for a particular class of Banach spaces, a "kernel trick" for non-linear classification.

1 Introduction

The theory of classical Support Vector Machines having attained an enviable apogee of mathematical completeness, empirical success and aesthetic coherence, efforts in the Machine Learning community have recently turned to possible extensions of its basic framework. Perhaps the prime restriction in the standard theory is the assumption that the training data (or its features) lie in a Hilbert space. This choice is natural as an initializing point: the geometry of Hilbert spaces is well understood, and the bilinearity of the inner product makes a thorough-going analysis possible. The simplicity of its structure also firmly marks its limitations — most data, for instance, do not come with any natural notion distance that can be induced from an inner-product. Hilbert spaces are also somewhat pedestrian: since all Euclidean spaces of the same basis cardinality are isometrically isomorphic [3], there is, in a sense, only one inner-product space.

The search for generalizations is then a search for algebraic/analytic structures more accommodating to larger classes of data, and more representative of complex distance relations. In the linear algebra hierarchy, the normed spaces and their complete cousins, the Banach spaces, inhabit the niche one place removed from the Hilbert spaces. These are spaces where lengths of vectors have been defined (and hence distances), but not angles. Research on large-margin classification in Banach spaces has already been initiated [1, 13], and even more generally in metric spaces [6, 10]; in an alternative direction [11] considers classification in Krein spaces, i.e. spaces with non-positive inner-products. Arguably, the Banach space setting still occupies the center-stage of scrutiny, for a number of reasons. First, vector space structures come ready-equipped with a number of convenient properties and objects; most importantly for classification problems, that of linear operators and functionals, and hence hyperplanes. Secondly, the introduction of a norm (and more generally, metric) allows the preservation of the notion of *margin*. These two objects: hyperplane and margin, might be construed as the minimal set of concepts required to construct a geometric generalization of classification in Hilbert spaces; if so assumed, the Banach space assumption then becomes the natural choice of minimal structure required to support these two notions. Finally, classification problems in cases where the data do not possess linear structure may still be attacked via normed-space ideas. For example, every metric space can be isometrically embedded into a Banach space: a number of constructions exist [6, 10]. Procedures developed for normed spaces may lead directly to algorithms for classification in the more general metric spaces.

What can replace the inner-product in non-Hilbert Banach spaces? G. Lumer in [9] introduced the notion of semi-inner product (s.i.p.) spaces: normed vector spaces with a type of inner-product satisfying many, but not all, of the axioms of a Hilbert inner-product. Crucially, every Banach space can be represented by a (not necessarily unique) s.i.p., a form with sufficient structure to carry over to the Banach space setting a number of Hilbert-space-type arguments. Indeed, many concepts seemingly unique only to Hilbert spaces find counterparts in normed spaces, via the semi-inner-product machinery. To give a sample selection: the Riesz Representation theorem and duality mappings, orthogonality relations, and generalizations of special concepts for Hilbert operators such as Hermiticity and numerical range.

This paper outlines a theory of large-margin binary classification in Banach spaces, where the central results are derived and couched in a semi-inner product formalism. In particular, we focus on a certain well-behaved class of Banach spaces: the uniformly smooth and uniformly convex spaces. Roughly speaking, these are structures possessing a type of Riesz representation theorem and include, for example, the L^p spaces, $1 < p < \infty$. In such spaces, the entire infrastructure for linear classification so well-studied in Hilbert spaces "goes over", more or less with the s.i.p. replacing the inner product. Indeed, we prove that the maximum-margin problem becomes well-posed, we establish a finite-dimensional linear Representer theorem, and show that the coefficients of the classifier are obtained through a convex (non-quadratic) optimization problem; remarkable facts given that the supporting s.i.p.'s are not bilinear in general.

For a special class of Banach spaces, L^{2p}, p an integer, a complete theory, extending to the case of *non-linear* classifiers, can be developed, as a parallel to kernel methods for Hilbert space classification. Here, moment functions replace kernel functions to give a version of the kernel trick. The available types of dependency relations becomes significantly broadened in this theory, utilising as it does $2p$-th order statistics instead of the second-order statistics of the inner-product.

We begin with a primer on s.i.p. spaces, collecting a number of results culled from the mathematical literature. Section 3 discusses the application of these results to classification in Banach space, by deriving, from the geometric s.i.p. point of view, two formulations of hyperplane classification, one an optimization over the learning domain, and the other in the dual of continuous linear functionals. The s.i.p. machinery shows that the optimization in the dual space can always be made a simple convex problem with *affine* constraints. A Representer theorem is then proved, demonstrating that it suffices to consider only linear functionals on the space spanned by the data; this result allows us to formulate a finite-dimensional convex program for the hyperplane coefficients using standard ideas from Lagrange optimization. Finally, the latter sections discuss how to obtain non-linear classifiers for the case L^{2p}, via moment functions, in an analagous generalization of Hilbert SVM theory.

2 Semi-Inner-Product Spaces

Let us collect a number of important results on semi-inner-products useful in the sequel. For simplicity of discourse, and with a view toward applications, we have not always provided the most general conditions for each statement: for optimal assumptions the references may be consulted.

Definition 1. *Let $(\mathcal{X}, \|\cdot\|)$ be a real Banach space. A semi-inner-product (s.i.p.) on $\mathcal{X}$ is a real function $\langle x, y\rangle$ on $\mathcal{X} \times \mathcal{X}$ with the properties*[1]

1. *(Linearity in second argument)* $\langle x, y_1 + y_2\rangle = \langle x, y_1\rangle + \langle x, y_2\rangle$
2. *(Homogeneity)* $\langle ax, y\rangle = \langle x, ay\rangle = a\langle x, y\rangle$
3. *(Norm-inducing)* $\langle x, x\rangle = \|x\|^2$
4. *(Cauchy-Schwartz)* $\langle x, y\rangle \le \|x\|\|y\|$

Semi-inner-products are not usually linear in their first argument, nor symmetric, unless the space is Hilbertian, in which case the s.i.p. coincides with the inner product. The Hahn-Banach theorem gives the existence of a s.i.p. for every Banach space, without providing any explicit description for the possible supporting s.i.p's, nor conditions under which the s.i.p. is unique. The special case of *smooth* Banach spaces (i.e. where the norm is Gâteaux differentiable) suffices to ensure uniqueness of the representation, as well as an explicit form for the s.i.p. in terms of the norm:

Theorem 1. *[4] A Banach space $\mathcal{X}$ has unique s.i.p. if and only if it is smooth, in which case*

$$\langle x, y\rangle = \lim_{\lambda\to 0} \frac{\|x+\lambda y\|^2 - \|x\|^2}{2\lambda} \qquad (1)$$

The above result is highly apposite for the calculation of s.i.p.'s, and shows that the semi-inner products are essentially directional derivatives of the square norm.

It will be desirable to consider classes of Banach spaces not only with differentiable norm, but which satisfy the following *uniform convexity property*: for each $\epsilon > 0$, there exists $\delta > 0$ such that $\|x+y\|/2 \le 1-\delta$ whenever $\|x-y\| > \epsilon$ for all x, y in the unit ball. Such spaces have several important characteristics; they are reflexive, and the infimum distance between closed convex sets C and a given point x_0 is actually achieved by

[1] The notation $\langle x, y\rangle$ for the s.i.p. must not be confused for the similar notation $\langle x^*, y\rangle \equiv x^*(y)$ sometimes employed for the evaluation of a linear functional x^* in the dual $\mathcal{X}^*$ at the point $y \in \mathcal{X}$. See, however, the generalized Riesz representation theorem of Theorem 2, where the two notations become somewhat unified.

some vector $c \in C$ [8]. We shall see that the uniform convexity assumption guarantees the existence and uniqueness of a maximum-margin hyperplane solution.

When a Banach space is both uniformly smooth and uniformly convex, one obtains a set of satisfying Hilbert-like duality properties:

Theorem 2. *Let $\mathcal{X}$ be a uniformly smooth and uniformly convex Banach space with s.i.p. $\langle\cdot,\cdot\rangle$ and dual $\mathcal{X}^*$. Then:*

i) [3] (General Riesz Representation) For each continuous linear functional $f \in \mathcal{X}^*$, there exists a unique vector $w \in \mathcal{X}$ such that $f(x) = \langle w, x\rangle$

ii) [2] The dual $\mathcal{X}^*$ is a uniformly smooth and uniformly convex Banach space supported by the semi-inner-product defined by $\langle f_{w_1}, f_{w_2}\rangle = \langle w_2, w_1\rangle$, where f_{w_i} is the linear functional associated with $w_i \in \mathcal{X}$.

Remark: We stress the alternation of positions of variables in the dual s.i.p.

Examples: All of the foregoing theorems are instructively illuminated in the following concrete situations.

1. $\mathcal{X} = L^p(\Omega,\mu)$. For $1 < p < \infty$, these Banach spaces are readily confirmed to be uniformly smooth and uniformly convex; this is not so for $p = 1$ or $p = \infty$. Let $\varphi_p : L^p(\Omega,\mu) \to L^q(\Omega,\mu)$ be defined[2] by $\varphi_p(x) = \frac{x^{\langle p-1\rangle}}{\|x\|_p^{p-2}}$, $\varphi_p(0) = 0$ through continuity, and $\frac{1}{p} + \frac{1}{q} = 1$. Then φ_p is a norm-preserving ($\|x\|_p = \|\varphi_p(x)\|_q$) bijection, with inverse φ_q, and $\langle x, y\rangle_p \equiv \int_\Omega \varphi_p(x) y \, d\mu$ defines the unique semi-inner-product on $\mathcal{X} = L^p(\Omega,\mu)$.

2. Stable Processes. Let $S(x)$ be a symmetric α-stable random process. The span of $(S(x))_{x\in\mathcal{X}}$ is a vector subspace $\mathcal{B}$ of the space of all stable random variables, and can be endowed with a norm: the spread $\sigma(S(x))$ (cf. [12]). This Banach space has an s.i.p. representation as follows. Define the *covariation* between $S(x_1)$ and $S(x_2)$ as

$$[S(x_1), S(x_2)] = \int_{S^1} s_1^{\langle\alpha-1\rangle} s_2 \, d\Gamma \qquad (2)$$

where s_1 and s_2 are (x,y) coordinates on the unit-circle, and Γ the spectral measure for the pair $(S(x_1), S(x_2))$. Then the (unique) semi-inner-product on $\mathcal{B}$ is defined by

$$\langle S_1, S_2\rangle = \frac{[S_1, S_2]}{\sigma_{S_1}^{\alpha-2}} \qquad (3)$$

for $S_1, S_2 \in \mathcal{B}$. The Gaussian case $\alpha = 2$ gives $\langle S_1, S_2\rangle = \frac{1}{2}\mathrm{Cov}(S_1, S_2)$.

One also has the peculiar representation, from a formula of Cambanis [12]:

$$\langle S_1, S_2\rangle = \frac{\sigma_{S_1}^2 E S_1^{\langle p-1\rangle} S_2}{E|S_1|^p}, \quad 1 < p < \alpha \qquad (4)$$

Semi-inner products induce a notion of orthogonality in normed linear spaces often helpful for geometric intuition: we define $x \perp y$ iff $\langle x, y\rangle = 0$. Note that because of asymmetry of the s.i.p., this notion of orthogonality is not usually symmetric. Seen in this light, the Riesz representation theorem of Theorem 3(i) is a generalization of the observation that in $\mathbb{R}^d$, a $(d-1)$-dimensional hyperplane passing through the origin is parameterized by a given normal vector w (in the s.i.p. sense) to the plane. It is not difficult to see from (1) that s.i.p. orthogonality coincides with the following notion of "minimum-distance" orthogonality in real normed linear spaces introduced by R. James [7, 5]: $x \perp y$ iff $\|x + \lambda y\| \geq \|x\|$ for all $\lambda \in \mathbb{R}$. It follows that many problems of best approximation in Banach spaces are naturally formulated in terms of semi-inner-products.

3 Hard-Margin Binary Classification in Banach Spaces

Our aim is to develop a semi-inner-product formulation of the maximal-hard-margin linear classification problem in Banach spaces. The advantage of such an approach over other developments, such as [1, 13], is that s.i.p. arguments emulating the Hilbert case become available. This allows us to go considerably farther in the development of a parallel theory. Moreover, the s.i.p. economically and clearly mediates between two equivalent formulations: one in the learning domain and one in the dual space. We shall see that in the general Banach space case, these two formulations are rather different, whereas in the Hilbert case they coincide since the dual of a Hilbert space is isometrically isomorphic to itself (i.e. in some sense self-dual).

Henceforth let us assume that the learning domain $\mathcal{X}$ is a uniformly smooth, uniformly convex Banach space with s.i.p. $\langle\cdot,\cdot\rangle$.

Lemma 1. *Given $w \in \mathcal{X}$, let $H = \{x \in \mathcal{X} : \langle w, x\rangle + b = 0\}$ be a hyperplane in $\mathcal{X}$. Then the distance between x_0 and H is $d = \inf_{x\in H} \|x_0 - x\| = \|w\|^{-1}|\langle w, x_0\rangle + b|$.*

[2] We employ the notation for the signed power function $a^{\langle b\rangle} = |a|^b \mathrm{sgn}(a)$ for $a \in R$ and $b > 0$, and the natural component-wise extension for a a vector or function.

Proof. This is simply Theorem 1 of [5], recast in the language of s.i.p.'s, via Theorem 3. ∎

Now let training points $\{x_i, y_i\}_{i=1}^m \in \mathcal{X}$ be given, where $y_i = \pm 1$. If the data are linearly separable, then there exists a (continuous) linear functional $f(x)$ and an offset $b \in \mathbb{R}$ such that $y_i(f(x_i)+b) > 0$, for all i. By Theorem 3, there exists a vector $w \in \mathcal{X}$ such that $f(x) = \langle w, x \rangle$. By rescaling w and b, using homogeneity of the s.i.p., and Lemma 1, we may assume without loss of generality that the point(s) closest to the hyperplane $H = \langle w, x \rangle + b$ satisfy $|\langle w, x_i \rangle + b| = 1$. Thus H may be placed in the canonical form $H = (w, b)$, with $y_i(\langle w, x_i \rangle + b) \geq 1$, for all i. With this form, it is also now immediate from Lemma 1 that the margin of the hyperplane is $\|w\|^{-1}$. We have then derived:

Data Domain Optimization for Maximum-Margin Banach Linear Classifier

$$\begin{aligned} &\inf_{w \in \mathcal{X}, b \in \mathbb{R}} \|w\|_{\mathcal{X}} \\ &s.t. \quad y_i(\langle w, x_i \rangle_{\mathcal{X}} + b) \geq 1 \end{aligned} \tag{5}$$

The classifier is given by $f(x) = \operatorname{sgn}(\langle w, x \rangle + b)$.

This of course is the usual hard-margin formulation in Hilbert spaces, with the inner product replaced by the semi-inner-product. Posing the problem in the dual space, through Theorem 3, we have

Dual Domain Optimization for Maximum-Margin Banach Linear Classifier

$$\begin{aligned} &\inf_{w^* \in \mathcal{X}^*, b \in \mathbb{R}} \|w^*\|_{\mathcal{X}^*} \\ &s.t. \quad y_i(\langle x_i^*, w^* \rangle_{\mathcal{X}^*} + b) \geq 1 \end{aligned} \tag{6}$$

It is instructive to compare the two problems (5) and (6). The key difference lies in the nature of the constraints: since the semi-inner product is linear in the second variable but generally non-linear in the first, one sees that the data-domain formulation gives rise to an optimization problem non-linear in its constraints, and in general *non-convex*, whereas the dual-form problem (6) gives rise to a *convex* optimization with *linear* constraints. Put another way, there exists an appropriate duality mapping (change of variables) of the non-convex problem (5) to the convex problem (6).

3.1 Existence and Uniqueness of the Solution

In general, there is no guarantee that the minimizer to the program (6) is unique: indeed it may not even exist. Simple counterexamples can be found in L^1, and the spaces of continuous functions, for instance. However, the imposition of *uniform convexity* does make the problem well-posed. Without loss of generality, assume $b = 0$. Define the sets $S_i = \{w^* : y_i(\langle x_i^*, w^* \rangle_{\mathcal{X}^*}) \geq 1\}$; these are closed, convex subsets of $\mathcal{X}^*$ for each i. Problem (6) can now be viewed as the task of finding the point in $\cap_i S_i$ closest to 0. As alluded to above, it is a standard fact from elementary functional analysis (c.f. [8]) that, given a point z in a uniformly convex Banach space $\mathcal{X}$, and a closed convex subset C of $\mathcal{X}$, there exists a unique point $c \in C$ such that $\|z - c\| = \inf_{c' \in C} \|z - c'\|$. This fact immediately produces

Theorem 3. *The solution to (5) and (6) exists and is unique, for uniformly smooth and uniformly convex Banach spaces $\mathcal{X}$.*

3.2 Form of the solution: A Linear Representer Theorem

The optimization problem (6) is posed, in general, in an infinite-dimensional space. However, since the number of data points is finite, one intuitively expects that the optimal solution should depend only on the metric relations between the data points; in other words, that one need only search in the space of functionals on the finite-dimensional space spanned by the data. Another way to put this is that the problem should not depend on the ambient space in which the data is embedded in. Such a theorem, in the Hilbert-space case, is known as the Representer Theorem. We shall prove this result now for uniformly smooth and uniformly convex Banach spaces.

The proof is constructed in two steps. The first, of interest in its own right, is the establishment of the necessity of a KKT-like condition for optimization problems with affine constraints in the generality of reflexive Banach spaces. The second step involves the computation of the associated Fréchet derivatives and an application of the semi-inner-product formalism to derive the required hyperplane representation.

For the moment, consider the more general setting of a reflexive Banach space $\mathcal{B}$, and a differentiable cost function $f : \mathcal{B} \to \mathbb{R}$. Suppose a solution to the linearly constrained problem

$$\begin{aligned} &\min_{x \in \mathcal{B}} f(x) \\ &\text{s.t.} \quad b_i + g_i(x) \geq 0, \quad \forall i = 1, \ldots, n \end{aligned} \tag{7}$$

is sought, where $\{g_i\}_{i=1}^n$ are continuous linear functionals in the dual $\mathcal{B}^*$, and $b_i \in \mathbb{R}$ shifting constants. Denoting by $D_x f \in \mathcal{B}^*$ the derivative of f at x, we have:

Theorem 4. *Any local minimum $x^\star \in \mathcal{B}$ to the optimization problem (7) satisfies $D_{x^\star} f = \sum_{i=1}^n \lambda_i g_i$, for some $\lambda_i \geq 0$.*

Proof. Critical to the proof is an application of a separating hyperplane theorem in the Banach space. Many versions exist; one which more than suffices for our purposes is:

Theorem 5. *[3] Let A and B be two disjoint nonempty convex sets in a real topological vector space X, with A open. Then there is a continuous linear functional $f^* \in X^*$, and $\alpha \in \mathbb{R}$ such that $f^*(a) < \alpha \leq f^*(b)$, $\forall a \in A, \forall b \in B$.*

Now we proceed via contradiction. Suppose that $D_{x^\star} f \notin C$, where C is the convex cone spanned by $\{g_1, \ldots, g_n\}$. Since C is closed in $\mathcal{B}^*$, there exists an open ball about $D_{x^\star} f$ not intersecting C; applying the separation theorem then gives an element $s^{**} \in \mathcal{B}^{**}$, and a real α such that $s^{**}(g_i) - \alpha \geq 0$ for all i, and $s^{**}(D_{x^\star} f) - \alpha < 0$. Since $s^{**}(0) = 0$, $\alpha \leq 0$; $s^{**}(C) \geq \alpha$ then implies $s^{**}(g_i) \geq 0$ and $s^{**}(D_{x^\star} f) < 0$. Reflexivity of the Banach space implies there exists an $s \in \mathcal{B}$ such that $s^{**}(x^*) = x^*(s)$, for all x^*, hence we have found an $s \in \mathcal{B}$ satisfying $g_i(s) \geq 0$ for all i and $(D_{x^\star} f)(s) < 0$.

Let $\delta > 0$, and consider the point $x^\star + \delta s$. Since $b_i + g_i(x^\star + \delta s) \geq 0$, $x^\star + \delta s$ is a feasible point. The differentiability of f implies

$$\begin{aligned} f(x^\star + \delta s) - f(x^\star) &= (D_{x^\star} f)(\delta s) + o(\|\delta s\|) \qquad (8) \\ &= \delta \cdot (D_{x^\star} f)(s) + o(|\delta|) \qquad (9) \end{aligned}$$

Thus there exists δ' sufficiently small such that for all $0 < \delta < \delta'$, $f(x^\star + \delta s) - f(x^\star) < 0$, contradicting the assumption that $x^\star$ is a local minimum. ∎

Return now to (6), and let $w^*_\star \in \mathcal{X}^*$ be its unique global minimizer. Let $w_\star \in \mathcal{B}$ be such that $\langle w_\star, \cdot \rangle_{\mathcal{X}} = w^*_\star$ (it exists uniquely by Theorem 3).

Theorem 6. *(A Representer Theorem). The maximum-margin separating hyperplane solving (5) admits the expansion*

$$w_\star = \sum_{i=1}^{m} \alpha_i x_i \qquad (10)$$

for some $\alpha_i \in \mathbb{R}$

Proof. Uniform smoothness and convexity of $\mathcal{B}$ implies the same for its dual $\mathcal{B}^*$. Let $f(w^*) = \|w^*\|^2$, differentiable because the space is smooth. Theorem 1 shows $2\langle w^*_\star, a^* \rangle_{\mathcal{X}^*} = (D_{w^*_\star} f)(a^*)$. Every uniformly convex Banach space is reflexive; Theorem 4 then states there exists $\lambda_i \geq 0$ such that

$$2\langle w^*_\star, a^* \rangle_{\mathcal{X}^*} - \sum_{i=1}^{m} \lambda_i y_i \langle x^*_i, a^* \rangle_{\mathcal{X}^*} = 0 \qquad (11)$$

$$\langle a, w_\star - \sum_{i=1}^{m} \frac{\lambda_i}{2} y_i x_i \rangle_{\mathcal{X}} = 0 \qquad (12)$$

The last line holds for every $a \in \mathcal{B}$, hence in particular for $a = w_\star - \sum_{i=1}^{m} \frac{\lambda_i}{2} y_i x_i$; this gives $\|w_\star - \sum_{i=1}^{m} \frac{\lambda_i}{2} y_i x_i\|^2 = 0$, the sought-after result with $\alpha_i = \frac{\lambda_i}{2} y_i$. ∎

For the special case of $\mathcal{X} = L^p(\Omega, \mu)$ spaces, we have established:

Corollary 1. *(Representer Theorem for L^p Classifier) The maximum-margin hyperplane $w^*_\star \in \mathcal{X}^*$ admits the expansion (with equality in the L^q sense)*

$$w^* = \left(\sum_{i=1}^{m} \alpha_i x_i \right)^{\langle 1/(q-1) \rangle} = \left(\sum_{i=1}^{m} \alpha_i x_i \right)^{\langle p-1 \rangle} \qquad (13)$$

Equivalently, if $w_\star \in \mathcal{X}$ is the unique representer for $w^*_\star$,

$$w_\star = \varphi_q(w^*_\star) = \frac{1}{C} \sum_i \alpha_i x_i \qquad (14)$$

for a real constant C; equality holding in the L^p sense.

3.3 Lagrange Dual S.i.p. Formulation

Direct substitution of the representation of Theorem 6 into (5) gives a finite-dimensional optimization problem, but one with non-convex constraints. Instead, we use Theorem 6 to assume that, without loss of generality, $\mathcal{X} = \text{span}\,\{x_1, \ldots, x_m\}$ and apply standard finite-dimensional convex optimization theory to the dual-space problem (6).

Form the Lagrange function to an equivalent convex problem of (6):

$$L(\lambda, w^*) = \frac{1}{2}\|w^*\|^2_{\mathcal{X}^*} + \sum_{i=1}^{m} \lambda_i (1 - y_i \left(b + \langle x^*_i, w^* \rangle_{\mathcal{X}^*} \right))$$

Fixing $\lambda_i \geq 0$, $\inf_{w^* \in \mathcal{X}^*} L(\lambda, w^*)$ is a convex problem with differentiable cost. It achieves its unique minimum when $\partial_{w^*} L = 0$ and $\partial_b L = 0$. The former has actually already been computed in (11), and implies $w = \sum_{i=1}^{m} \lambda_i y_i x_i$; the latter derivative implies $\sum_{i=1}^{m} \lambda_i y_i = 0$. The Lagrange dual function, using the Riesz Theorem to revert back to data variables becomes

$$\begin{aligned} L(\lambda) &= \frac{1}{2}\|w\|^2_{\mathcal{X}} + \sum_{i=1}^{m} \lambda_i (1 - y_i \langle w, x_i \rangle_{\mathcal{X}}) \qquad (15) \\ &= \frac{1}{2}\|w\|^2_{\mathcal{X}} + \sum_{i=1}^{m} \lambda_i - \langle w, \sum_{i=1}^{m} \lambda_i y_i x_i \rangle_{\mathcal{X}} \qquad (16) \\ &= -\frac{1}{2}\| \sum_{i=1}^{m} \lambda_i y_i x_i \|^2_{\mathcal{X}} + \sum_{i=1}^{m} \lambda_i \qquad (17) \end{aligned}$$

The convex dual optimization problem now has the structure:

Lagrange-Dual Optimization for Maximum-Margin Banach Linear Classifier

$$\max_{\lambda \in \mathbb{R}^m} -\frac{1}{2}\|\sum_{i=1}^m \lambda_i y_i x_i\|_{\mathcal{X}}^2 + \sum_{i=1}^m \lambda_i \tag{18}$$
$$\text{s.t.} \quad \lambda_i \geq 0, \quad \sum_{i=1}^m \lambda_i y_i = 0$$

Since the primal dual-space problem (6) is convex with affine constraints, strong duality is achieved through the Lagrange dual, by standard theorems in finite-dimensional convex analysis [2]. A solution to (18) then gives the solution also to the problem (5), with the large-margin classifier attaining the final form $f(x) = \mathrm{sgn}(\langle \sum_{i=1}^m \lambda_i y_i x_i, x\rangle_{\mathcal{X}} + b)$. The offset may be computed via the standard KKT condition that the product of dual variables and constraints must vanish [2], i.e. for any i for which $\lambda_i > 0$, $b = 1/y_i - \langle \sum_{j=1}^m \lambda_j y_j x_j, x_i\rangle_{\mathcal{X}}$. As in the Hilbert case, certain identities hold true: e.g. by summing over the aforementioned condition, the property $\sum_i \lambda_i = \|w\|^2 = 1/(\mathrm{margin})^2$ holds for the optimal hyperplane.

One observes then, that the usual SVM Lagrange dual optimization for Hilbert spaces generalizes naturally and directly to the Banach space case; the crucial difference being that the resulting classifier inhabits the structure of a semi-inner-product rather than an inner product, and hence exhibits a non-linear dependence with respect to the dual coefficients λ. Figure 1 displays a simple configuration of three labelled points, and the resulting large-margin classifiers computed with respect to the p-norms $p = 1.5, 2, 3$ and 4.

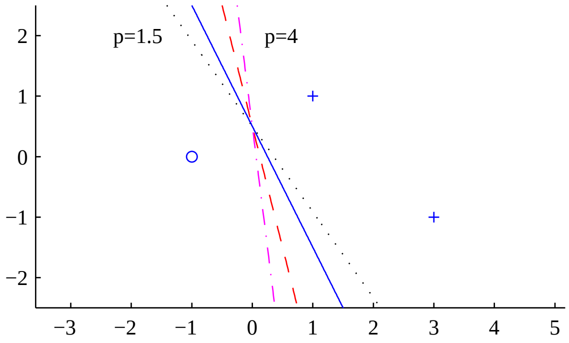

Figure 1: Maximum-margin separating hyperplanes in $(\mathbb{R}^2, \|\cdot\|_p)$.

3.4 Non-Linear Classifiers in L^p, $p \in 2\mathbb{Z}^+$

One of the most arresting ideas in standard SVM theory consists in the *kernel trick*: the procedure where inner-products in a learning domain $\mathcal{X}$ are replaced by bivariate functions $K(\cdot,\cdot)$ having the effect of implicitly mapping a problem into a (usually) higher-dimensional Hilbert space $\mathcal{H}$, through a feature map $\Phi : \mathcal{X} \to \mathcal{H}$. The "trick" consists in selecting a kernel whose calculation does not involve explicit knowledge of the map Φ, nor an inner-product evaluation in $\mathcal{H}$. In this way, classification may be performed in the high-dimensional, even infinite-dimensional feature space $\mathcal{H}$ without incurring the expected additional cost of a dimensionality increase.

Having developed an s.i.p. formalism for the Banach space binary classification problem, we are led immediately to the question of whether a similar "kernel trick" is available for semi-inner-products. An initial idea is, in emulation of the Hilbert counterparts, to define bivariate s.i.p. kernels. Two crucial aspects, however, become profound obstacles to the establishment of a similar "kernel" theory for Banach spaces along this route: (1) lack of bilinearity of the s.i.p. prevents the classifier $\langle \sum_{i=1}^m \lambda_i y_i x_i, x\rangle$ from being written as a function of s.i.p.'s between data points and test points, and (2) the s.i.p. does not prescribe the structure of a Banach space in as total a way as an inner product; for example, an inner product defined on a vector basis extends uniquely to the whole space: this is false for s.i.p.'s.

While there may thus appear to be little hope of establishing a *general* kernel theory for Banach spaces, nevertheless there is one special class of Banach spaces which appear highly amenable to a type of kernel theory: the L^p spaces for even integer p. We shall see that a certain type of $p-variate$ multi-linear moment function can take the place of the bivariate kernel function for Hilbert spaces; this theory, which we begin to delineate here, coupled with the linear theory of the previous section combine to give classification tools as powerful as the ones for Hilbert spaces.

Reconsider the $\mathcal{X} = L^p(\Omega,\mu)$ spaces, for even integers p. It will be more convenient to use the following equivalent optimization program to (18) for these spaces:

$$\max_{\lambda \in \mathbb{R}^m} -\frac{1}{p}\int_\Omega \left(\sum_{i=1}^m \lambda_i y_i x_i\right)^p d\mu + \sum_{i=1}^m \lambda_i \tag{19}$$
$$\text{s.t.} \quad \lambda_i \geq 0, \quad \sum_{i=1}^m \lambda_i y_i = 0$$

with the classifier now having type $f(x) = \mathrm{sgn}(b + \int_\Omega (\sum_{i=1}^m \lambda_i y_i x_i)^{p-1} x\, du)$.

Now we apply the non-linear extension. Let the data-points x_i belong to an abstract set $\mathcal{M}$, and pre-process via a map $\Phi : \mathcal{M} \to \mathcal{X}$ from the data-domain into an L^p feature space.

The quantities $\int_\Omega \left(\sum_{i=1}^m \lambda_i y_i \Phi(x_i)\right)^p d\mu$ and $\int_\Omega (\sum_{i=1}^m \lambda_i y_i \Phi(x_i))^{p-1} x \, du)$ may be respectively written as p-th and $(p-1)$-th order polynomials in $\lambda_i y_i$, with coefficients of the form $\int_\Omega \Phi(x_{i_1}) \cdots \Phi(x_{i_p}) \, d\mu$ — p-th order moment functions in the feature space. The optimization may then proceed without explicit knowledge of Φ, but simply via moment functions $M(x_1, \ldots, x_p) = \int_\Omega \Phi(x_1) \cdots \Phi(x_p) \, d\mu$. The choice of different moment functions implicitly provides a selection of non-linear map into an L^p space, in absolute analogy to the L^2 case, restricting the search to a small hypothesis space of possible non-linear classifiers induced by M. The even-order L^p non-linear classifier, given a moment function $M(x_1, \ldots, x_p)$, is then the solution of the convex program:

Optimization for Maximum-Margin L^p Moment Classifier

$$-\frac{1}{p} \max_{\lambda_i} \sum_{(i_1,\ldots,i_p)=1}^{m} \lambda_{i_1} y_{i_1} \cdots \lambda_{i_p} y_{i_p} M(x_{i_1}, \ldots, x_{i_p}) + \sum_i \lambda_i$$

$$\text{s.t.} \quad \lambda_i \geq 0, \quad \sum_{i=1}^m \lambda_i y_i = 0 \tag{20}$$

resulting in the non-linear classifier $f(x) = \operatorname{sgn}(b + \sum_{(i_1,\ldots,i_{p-1})=1}^m \lambda_{i_1} y_{i_1} \cdots \lambda_{i_{p-1}} y_{i_{p-1}} M(x_{i_1}, \ldots, x_{i_{p-1}}, x))$. The offset b is once more computed with $b = 1/y_k - \sum_{(i_1,\ldots,i_{p-1})=1}^m \lambda_{i_1} y_{i_1} \cdots \lambda_{i_{p-1}} y_{i_{p-1}} M(x_{i_1}, \ldots, x_{i_{p-1}}, x_k))$ for any k satisfying $\lambda_k > 0$.

3.5 Construction of Moment Functions

To specify a moment function M in an L^p space is to give not only its semi-inner product $\langle x, y \rangle$, but in fact the general p-th order statistic. Moment functions therefore contain significantly more information than the s.i.p. representation of the Banach space, specifying the L^p structure in a way similar to kernel functions for L^2. Such moment functions will satisfy certain properties: multi-linearity in the feature space, exchangeability in its p variables, positivity ($M(x, \ldots, x) \geq 0$), as well as various Hölder-like inequalities; to illustrate in the case $p = 4$: $M^4(x, x, x, y) \leq M^3(x, x, x, x) M(y, y, y, y)$ and $M^4(x, x, y, y) \leq M^2(x, x, x, x) M^2(y, y, y, y)$.

A rich and general source of feature spaces consist of spaces of random variables — as a familiar example, Gaussian feature maps Φ. Let $k(x, y)$ be a positive-definite kernel on the data-domain $\mathcal{M} = \mathbb{R}^d$, and $G(x)$ the associated zero-mean Gaussian random process on $\mathbb{R}^d$ with covariance k; define the feature map Φ by $x \to G(x)$. The higher-order moments in this case are easily calculable in terms of the second-order structure:

$$\begin{aligned} M(x_1, \ldots, x_p) &= E(G(x_1) \cdots G(x_p)) \\ &= \sum_\pi (k(x_i, x_j) \cdots k(x_w, x_z)) \end{aligned} \tag{21}$$

over permutations $\pi \in S_p$, for a total of $(p-1)!/(2^{p/2-1}(p/2-1)!)$ terms, each a product of $p/2$ kernel terms. For example, $p = 4$ gives $M(x_1, \ldots, x_4) = k(x_1, x_2)k(x_3, x_4) + k(x_1, x_3)k(x_2, x_4) + k(x_1, x_4)k(x_2, x_3)$.

Observe $p = 2$ gives the usual SVM classifier with kernel k. The interpretation for $p > 2$ is that one uses the same feature map Φ into the space of Gaussian random variables, but with the geometry in that space induced by the p-norm rather than the 2-norm. In general, one class of feature spaces arise from the probability measure P of a random process $\Phi(x)$ on $\mathbb{R}^d$; the moment functions M are then functions of the measure P. However, by using statistics of order higher than 2, we allow dependency structures not available to Gaussian processes. In the Hilbert-space case, every inner-product kernel can be achieved by a Gaussian feature map — this follows from the spectral theorem. Not so with the non-linear Banach classifiers. Indeed, one may imagine different feature mappings Φ_1 and Φ_2 into spaces of random variables sharing the same second-order kernel structure, but with different p-order statistics; here the Hilbert classifiers agree, the Banach classifiers differ.

Many other moment functions can be generated from kernel functions by using feature maps of the type $\Phi : x \to f(G(x)$, for $f : \mathbb{R} \to \mathbb{R}$. For example, with $f = \exp(\cdot)$ one obtains a log-normal random process parameterized by kernel k. Using characteristic functions, trite calculations show that

$$M(x_1, \ldots, x_p) = \exp\left(\sum_{i=1}^p \sum_{j=1}^p k(x_i, x_j)\right) \tag{22}$$

is an admissible moment function for any kernel k.

Other easy facts concerning the combination properties of moment functions can be derived with basic probability, assuming finite-measure feature spaces. We shall include only a brief listing here of the uncountable variations. Let Φ_1 and Φ_2 be two independent L^p random processes on $\mathbb{R}^d$ with moment functions M_1 and M_2. Then the product $\Phi_1 \cdot \Phi_2$ defines a feature map with moment function $M = M_1 \cdot M_2$, and is hence admissible. If Φ is the map taking each x to a constant $m(x) \equiv m_x : \Omega \to \mathbb{R}$, then $M(x_1, \ldots, x_p) = \prod_{i=1}^p m(x_i)$ is an admissible moment function. Let M_1 and M_2 be two Gaussian 4-th order moment functions generated from kernels k_1 and k_2.

Then

$$\begin{aligned} M_1(x,y,z,w) &+ k_2(x,z)k_1(y,w) + k_2(y,z)k_1(x,w) \\ &+ k_2(x,y)k_1(z,w) + k_2(w,z)k_1(x,y) \\ &+ k_2(w,x)k_1(y,z) + k_2(y,w)k_1(x,z) \\ &+ M_2(x,y,z,w) \end{aligned}$$

is admissible. This last result is one 4-th order generalization of the second-order fact that kernels form a cone.

4 Discussion

We have developed a semi-inner-product formulation of the binary classification problem in Banach spaces. The main message might be said to be that all of Hilbert linear classification theory carries over neatly to the case of Banach spaces (at least well-behaved ones). The resulting optimization programs are no longer quadratic, but remain convex. Finally, even *kernel* theory has its appropriate generalization in the spaces L^{2p}, where moment functions replace kernel functions. In our presentation of this non-linear classification framework, we have made no claims to having constructed the complete story, but merely illustrated by way of calculations that an analogous theory and practice to the Hilbert case exists, and should be further studied. Certain apposite questions still beg to be answered. What characterizations exist for moment functions, in the vein of Mercer's theorem for kernels? Is there a way to produce spectral decompositions of the moment tensors? Is there a corresponding notion of Reproducing Kernel Banach Space (RKBS), and, if so, how does one construct an RKBS given an s.i.p. (or stronger, a moment function)? What generalization error bounds can be established for Banach classifiers, and how does one choose the moment functions relative to the data? Can non-linear classification be extended to general L^p spaces, for fractional values of p, for instance, by approximating the semi-inner product with polynomials of moment functions?

These and similar points will be addressed in future work.

References

[1] K. Bennett, E. Bredensteiner, *Duality and Geometry in SVM classifiers*, Proceedings of the Seventeenth International Conference on Machine Learning, pp. 57–64, (2000).

[2] S. Boyd, L. Vandenberghe, *Convex Optimization.* Cambridge University Press, (2004).

[3] J. Conway, *A Course in Functional Analysis.* Springer Science, (1990).

[4] S. Dragomir, *Semi-inner products and Applications.* Nova Science, Hauppauge, New York, (2004).

[5] J. Giles, *Classes of Semi-Inner-Product Spaces*, Transactions of the American Mathematical Society, Vol. 129, No. 3, pp. 436–446, (1967).

[6] M. Hein, O. Bousquet, B. Schölkopf, *Maximal Margin Classification for Metric Spaces*, Journal of Computer System Sciences, Vol. 71, No. 3, pp. 333—359, (2005).

[7] R. James, *Orthogonality and Linear Functionals in Normed Linear Spaces*, Transactions of the American Mathematical Society, Vol. 61, No. 2, pp. 265–292, (1947).

[8] P. Lax, *Functional Analysis.* John Wiley & Sons, (2002).

[9] G. Lumer, *Semi-Inner-Product Spaces*, Transactions of the American Mathematical Society, Vol. 100, No. 1, pp. 29–43, (1961).

[10] U. Luxburg, O. Bousquet, *Distance-Based Classification with Lipschitz Functions*, Journal of Machine Learning Research, Vol. 5, pp. 669–695, (2004).

[11] C. Ong, X. Mary, S. Canu, A. Smola, *Learning with Non-Positive Kernels*, Proceedings of the International Conference on Machine Learning, (2004).

[12] G. Samorodnitsky, M. Taqqu, *Stable Non-Gaussian Random processes.* Chapman & Hall, New York, (1994).

[13] D. Zhou, B. Xiao, H. Zhou, R. Dai, *Global Geometry of SVM Classifiers*, Technical Report 30-5-02, Institute of Automation, Chinese Academy of Sciences, (2002).