The regularized least squares algorithm and the problem of learning halfspaces

Ha Quang Minh

Institute for Theoretical Biology, Humboldt University of Berlin, Invalidenstrasse 43, Berlin D-10115, Germany

ARTICLE INFO

Article history:
Received 20 February 2009
Received in revised form 30 November 2009
Accepted 17 January 2011
Available online 20 January 2011
Communicated by P.M.B. Vitányi

Keywords:
Computational complexity
Analysis of algorithms
Computational learning theory
Halfspaces
Regularized least squares algorithm
Kernel

ABSTRACT

We provide sample complexity of the problem of learning halfspaces with monotonic noise, using the regularized least squares algorithm in the reproducing kernel Hilbert spaces (RKHS) framework.

1. Introduction

The theory of reproducing kernel Hilbert spaces (RKHS) has recently emerged as a powerful framework for the problem of learning from data, both from algorithmic and theoretical perspectives (we refer to, for example, [25,24, 20]). One well-known and very general algorithm in the RKHS framework is regularized least squares (see for example [25,6,16]), with the regularization term being an RKHS norm. Although originally developed as a regression algorithm, regularized least squares has been successfully applied in various practical classification problems (some examples can be found in [18,19,17,1]). This paper is a theoretical study of regularized least squares in RKHS for a special and very important binary classification problem, namely the problem of learning halfspaces. This problem has been studied intensively in the computational learning theory literature (see [7,9–12,4,2,8,5,21]). In this paper, we show that regularized least squares in RKHS can be applied to learn successfully halfspaces with monotonic noise. It is our hope that the results contained herein contribute to the theoretical understanding of this algorithm with regard to the binary classification problem in general. Because of the generality of the RKHS framework, we also hope that algorithms such as this can be applied to solve other problems of interest in computational learning theory.

1.1. Organization of the paper

We will first give a review of the definition of the binary classification problem in Section 2. The problem of learning halfspaces, along with previous works, is reviewed in Section 3. Our main results, which are concerned with sample complexity of halfspace learning, are stated in Section 5 and proved in Section 6.

2. The binary classification problem

Let the input space $X$ be a closed subset of $\mathbb{R}^n$ (more generally, any space on which a probability measure can be defined), such that the points in $X$ belong to two distinct categories, with labels $Y = \{-1, 1\}$. Suppose that there is an unknown probability measure $\rho$ on $Z = X \times Y$ which admits the decomposition
\( \rho(x, y) = \rho_X(x) \rho(y|x) \).

A binary classifier is a Boolean function \( f : X \to \{-1, 1\} \) that assigns either the \(-1\) or \(1\) label to each point \( x \in X \).

Consider the regression function [6], defined by

\[
f_\rho(x) = \int y \, d\rho(y|x) = P(y = 1|x) - P(y = -1|x) \tag{1}
\]

which satisfies

\[
P(y = \text{sgn}(f_\rho(x))) \geq P(y \neq \text{sgn}(f_\rho(x))|x). \tag{2}
\]

The optimal binary classifier, the Bayes classifier, is therefore

\[
\text{sgn} f_\rho(x) = \begin{cases} 1, & \text{if } P(y = 1|x) \geq P(y = -1|x), \\ -1, & \text{if } P(y = 1|x) < P(y = -1|x). \end{cases} \tag{3}
\]

Let \( L_2^\rho(X) \) denote the Hilbert space of squares integrable functions on \( X \), with norm denoted by \( \| \cdot \|_\rho \). Let \( f : X \to \mathbb{R} \) be a real-valued function, then it induces the binary classifier \( \text{sgn} f(x) : X \to \{-1, 1\} \), where \( \text{sgn}(f(x)) = \text{sgn}(f(x)) \).

The error of this classifier with respect to the Bayes classifier is then

\[
\rho_X(X_f) = \frac{1}{4} \| \text{sgn}(f) - \text{sgn}(f_\rho) \|_{\rho}^2, \tag{4}
\]

where \( X_f = \{ x \in X : \text{sgn}(f(x)) \neq \text{sgn}(f_\rho(x)) \} \). We do not know the ideal minimizer \( f_\rho \), since \( \rho \) is unknown, but we have access to random examples from \( X \times Y \) sampled according to \( \rho \). Let \( z = (x', y')_{i=1}^m \) be a finite random sample of size \( m \), drawn independently according to \( \rho \). The task of the binary classification problem is then to construct binary classifiers \( \text{sgn}(f_2) \) that approximate \( \text{sgn}(f_\rho) \) in the \( \| \cdot \|_{\rho} \) norm, using the finite sample \( z \).

It is required that

\[
\lim_{m \to \infty} \frac{1}{4} \| \text{sgn}(f_2) - \text{sgn}(f_\rho) \|_{\rho}^2 = 0, \tag{5}
\]

with high probability.

3. The problem of learning halfspaces

The problem of learning halfspaces is the most special case of the binary classification problem and one of the first studied by machine learning researchers. Let \( n \in \mathbb{N} \) be fixed. For a fixed \( w_0 \in \mathbb{R}^n \), there corresponds a unique hyperplane \( H_{w_0} = \{ x \in \mathbb{R}^n : \langle w_0, x \rangle = 0 \} \). Consider the binary classification problem where the two classes are linearly separated, with \( H_{w_0} \) as separating boundary. The learning target is the Boolean function \( f(x) = \text{sgn}(\langle w_0, x \rangle) \), called the halfspace associated with \( H_{w_0} \).

3.1. Halfspaces with monotonic noise

Let \( S^{n-1} = \{ x \in \mathbb{R}^n : \| x \| = 1 \} \) be the \( n \)-dimensional unit sphere with surface measure \( |S^{n-1}| = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \). Consider the case where the data lie on \( X = S^{n-1} \) and the labels are corrupted by noise. By symmetry on the sphere, there is no loss of generality by supposing that \( w_0 = (0, \ldots, 0, 1) \).

Let \( \hat{\eta} : [0, 1] \to [0, 1] \) be a nonincreasing function and suppose that

\[
P(y = 1 | x \in S^{n-1}) = 1 - \hat{\eta}(|x_0|),
\]

\[
P(y = -1 | x \in S^{n-1}) = \hat{\eta}(|x_0|),
\]

\[
P(y = -1 | x \in S^{n-1}) = 1 - \hat{\eta}(|x_0|). \tag{6}
\]

By definition, the regression function is given by

\[
f_\rho(x) = \begin{cases} 1 - 2\hat{\eta}(|x_0|), & \text{if } x \in S^{n-1} (x_0 > 0), \\ -(1 - 2\hat{\eta}(|x_0|)), & \text{if } x \in S^{n-1} (x_0 < 0). \end{cases} \tag{7}
\]

If \( \| \hat{\eta} \|_\infty < \frac{1}{2} \), then clearly

\[
\text{sgn}(f_\rho) = \text{sgn}(\langle w_0, x \rangle),
\]

and the halfspace \( \langle w_0, x \rangle \) is the optimal Bayes classifier. The problem of learning the halfspace \( \langle w_0, x \rangle \) is then said to have monotonic noise \( \hat{\eta} \).

Assume that the overall probability of misclassification is \( \eta \) for some fixed \( 0 \leq \eta < \frac{1}{2} \), that is

\[
\eta = \frac{1}{|S^{n-1}|} \int \hat{\eta}(x) \, dS^{n-1}(x). \tag{9}
\]

For \( \hat{\eta}(t) \equiv \eta, 0 \leq \eta < 1/2 \), the problem is said to have random classification noise \( \eta \). If \( \hat{\eta}(t) \equiv 0 \), the problem is said to be noise-free.

Remark 1. If \( \| \hat{\eta} \|_\infty > \frac{1}{2} \), then \( \text{sgn}(\langle w_0, x \rangle) \) is no longer the optimal Bayes classifier.

3.2. Previous work

The problem of learning halfspaces, where \( \rho_X \) is the uniform distribution on \( S^{n-1} \), has been studied extensively in computational learning theory. In the noise-free case, it was shown in [11, 12] that halfspaces, with respect to the uniform distribution on \( S^{n-1} \), are learnable to accuracy \( \epsilon \) with sample complexity \( \Theta(n^{2\epsilon - 1}) \). The Perceptron algorithm was shown in [4] to have sample complexity \( \tilde{O}(\frac{n}{\epsilon}) \) and running time \( O(n^2) \).

The algorithm in [10] has sample complexity \( \tilde{O}(\frac{n^2}{\epsilon^2}) \) and running time \( \tilde{O}(\frac{n^2}{\epsilon^2} + n^3) \), where \( \tilde{O}(a) = O(a \log^a a) \) for some \( p > 0 \).

Random classification noise was introduced in [2]. In the statistical query learning framework, [8] gave an algorithm with sample complexity \( \tilde{O}\left(\frac{n^2}{\epsilon^2(1-2\eta)^2}\right) \) and running time \( \tilde{O}\left(\frac{n^2}{\epsilon^2(1-2\eta)^2}\right) \). Monotonic noise was introduced in [5]. Under the monotonic noise assumption, the Average algorithm of [21] has sample complexity \( \tilde{O}\left(\frac{n}{\epsilon^2(1-2\eta)^2}\right) \) and running time \( \tilde{O}\left(\frac{n^2}{\epsilon^2(1-2\eta)^2}\right) \).

4. RKHS and the regularized least squares algorithm

In this section, we briefly review RKHS. The general theory of reproducing kernel Hilbert spaces was developed
by Aronszajn [3]. Let \( X \) be an arbitrary nonempty set. Let \( K : X \times X \to \mathbb{R} \) be a symmetric function satisfying: for any finite set of points \( \{x_i\}_{i=1}^N \) in \( X \) and real numbers \( \{a_i\}_{i=1}^N \),
\[
\sum_{i,j=1}^N a_i a_j K(x_i, x_j) \geq 0.
\]
(10)

\( K \) is said to be a **positive definite kernel** on \( X \). There exists a unique Hilbert space \( \mathcal{H}_K \) of functions on \( X \) satisfying:

1. \( K_x \in \mathcal{H}_K \) for all \( x \in X \), where \( K_x(t) = K(x, t) \);
2. \( \text{span}\{K_x\}_{x \in X} \) is dense in \( \mathcal{H}_K \);
3. the inner product \( \langle \cdot, \cdot \rangle_K \) of \( \mathcal{H}_K \) satisfies:

\[
f(x) = \langle f, K \rangle_K \quad \forall f \in \mathcal{H}_K \quad \text{(reproducing property)}.
\]

\( \mathcal{H}_K \) is called the Reproducing Kernel Hilbert Spaces with reproducing kernel \( K \). The regularized least squares algorithm of learning theory (see [16] and references therein) attempts to solve both problems of least squares regression and binary classification by the following minimization procedure.

**Algorithm 1.** Let \( K : X \times X \to \mathbb{R} \) denote a positive definite kernel. Let \( \mathcal{H}_K \) be the corresponding RKHS, with norm \( \|f\|_K \). For each \( \lambda > 0 \), let

\[
f_{f, \lambda} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}.
\]

For the binary classification problem, \( \text{sgn}(f_{f, \lambda}) \) is taken to be the empirical version of \( \text{sgn}(f_{\rho}) \), which approximates \( \text{sgn}(f_{\rho}) \) in the \( \| \cdot \|_\rho \) norm.

The solution \( f_{f, \lambda} \) is given by

\[
f_{f, \lambda} = \sum_{i=1}^m a_i K(x_i, \cdot),
\]

(12)

where \( a = (a_1, \ldots, a_m)^T \) is the unique solution of the system of linear equations

\[
(K[x] + \lambda m I)a = y,
\]

(13)

with \( y = (y_1, \ldots, y_m)^T \), \( K[x] \) the \( m \times m \) matrix whose \( (i, j) \)-entry is \( K(x_i, x_j) \), with \( I \) being the identity matrix.

**Remark 2.** The solution \( f_{f, \lambda} \) is obtained by solving a system of linear equations of size \( m \times m \). Thus constructing \( f_{f, \lambda} \) takes time \( O(m^3) \).

5. **Main results of the paper**

**Theorem 1.** Let \( \rho \) be a probability distribution on \( S^{n-1} \times \{-1, 1\} \) satisfying condition (6), with \( \rho_X \) being the uniform distribution on \( S^{n-1} \). Let \( z \) be randomly sampled according to \( \rho \). Then for \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),

\[
\rho_X(X f_{f, \lambda}) = \frac{1}{4} \| \text{sgn}(f_{f, \lambda}) - \text{sgn}(w_0, x) \|_\rho^2 \leq \frac{14 \sqrt{n} \log \frac{2}{\delta}}{\sqrt{m(1 - 2\eta)}}
\]

with \( \lambda = \frac{1}{\eta} \), provided that

\[
m \geq \frac{573 n (\log \frac{2}{\delta})^2}{(1 - 2\eta)^2}.
\]

**Remark 3.** It follows that for each \( \epsilon > 0 \), we have \( \rho_X(X f_{f, \lambda}) \leq \frac{1}{\epsilon} \) if

\[
m \geq \frac{196 n (\log \frac{2}{\delta})^2}{\epsilon^2 (1 - 2\eta)^2} = \Theta \left( \frac{n}{\epsilon^2 (1 - 2\eta)^2} \right).
\]

For \( 0 < \epsilon < 1/2 \), condition (15) is then automatically satisfied. The running time for constructing \( f_{f, \lambda} \) is then \( O \left( \frac{n^3}{\epsilon^6 (1 - 2\eta)^6} \right) \).

**Remark 4.** Thus for the problem of learning halfspaces with monotonic noise, where \( \rho_X \) is the uniform distribution on \( S^{n-1} \), our analysis of Algorithm 1 with \( K(t, x) = (t, x) \) gives the same sample complexity as the Average algorithm of [21], up to an absolute multiplicative constant, but with worse running time. This should not be unexpected, however, since the RKHS least squares method is a very general algorithm, whereas the Average algorithm is specifically for this particular problem.

**Theorem 2.** Let \( \rho \) be a probability distribution on \( S^{n-1} \times \{-1, 1\} \) satisfying (6), with \( \rho_X \) being the uniform distribution on \( S^{n-1} \). Let \( z \) be randomly sampled according to \( \rho \). Then for \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),

\[
\frac{1}{4} \left\| \text{sgn} \left( \frac{1}{m} \sum_{i=1}^m y_i x_i \right) - \text{sgn}(w_0, x) \right\|_\rho^2 \leq \frac{5 \sqrt{n} \log \frac{2}{\delta}}{\sqrt{m(1 - 2\eta)}},
\]

(17)

provided that \( m \geq \frac{64 n (\log \frac{2}{\delta})^2}{(1 - 2\eta)^2} \).

**Remark 5.** This is readily recognizable as a form of the Average algorithm of [21]. We will show that it also follows from our RKHS framework.

6. **Proofs of main results**

6.1. **Proof methodology**

There is an approach in the literature (we refer to [26] and references therein) to compute \( \rho_X(X f_{f, \lambda}) = \frac{1}{4} \| \text{sgn}(f_{f, \lambda}) - \text{sgn}(f_{\rho}) \|_\rho^2 \) from \( \| f_{f, \lambda} - f_{\rho} \|_\rho \), given some appropriate assumptions, such as Tsybakov’s noise condition [23]. However, for our setting, given that the kernels we use are continuous and specifically for this particular problem.

\[
\rho_X(X f_{f, \lambda}) = \frac{1}{4} \| \text{sgn}(f_{f, \lambda}) - \text{sgn}(w_0, x) \|_\rho^2 \leq \frac{14 \sqrt{n} \log \frac{2}{\delta}}{\sqrt{m(1 - 2\eta)}}
\]

with \( \lambda = \frac{1}{\eta} \), provided that

\[
m \geq \frac{573 n (\log \frac{2}{\delta})^2}{(1 - 2\eta)^2}.
\]
Lemma 1. Let \[\hat{\eta}(t)\] and \(\eta\) be as above, then

\[
\int_0^1 \hat{\eta}(t)(1 - t^2)^{\frac{n-3}{2}} dt \leq \int_0^1 \eta(t)(1 - t^2)^{\frac{n-3}{2}} dt. \tag{20}
\]

Proof. By definition, we have

\[
\eta = \frac{2|S^{n-2}|}{|S^{n-1}|} \int_0^1 \hat{\eta}(t)(1 - t^2)^{\frac{n-3}{2}} dt.
\]

Since \(|S^{n-1}| = 2|S^{n-2}| \int_0^1 (1 - t^2)^{\frac{n-3}{2}} dt\), it follows that

\[
\int_0^1 \hat{\eta}(t)(1 - t^2)^{\frac{n-3}{2}} dt = \int_0^1 \eta(t)(1 - t^2)^{\frac{n-3}{2}} dt = \frac{|S^{n-1}|}{2|S^{n-2}|} \eta.
\]

From the definition of \(\eta\) and the fact that \(\hat{\eta}\) is nonincreasing, there must be a point \(t_0\) in (0, 1) such that \(\hat{\eta}(t) \geq \eta\) for \(0 \leq t < t_0\) and \(\hat{\eta}(t) \leq \eta\) for \(t_0 < t \leq 1\). Thus

\[
\int_0^1 t(\hat{\eta}(t) - \eta)(1 - t^2)^{\frac{n-3}{2}} dt
\]

and

\[
\int_0^1 t_0(\hat{\eta}(t) - \eta)(1 - t^2)^{\frac{n-3}{2}} dt
\]

as required. \(\square\)

Proposition 1. Let \(x\) be randomly sampled according to \(\rho\) satisfying \(|y| \leq M\) almost surely. Then for any \(0 < \delta < 1\), with probability at least \(1 - \delta\),

\[
\|f_{x,\lambda} - f_{\lambda}\| \leq \frac{6\delta M \log^2 2}{\sqrt{m\lambda}}. \tag{18}
\]

6.2. The regularized least squares algorithm

We note that for \(K(x, t) = \langle x, t \rangle\) on \(S^{n-1} \times S^{n-1}\), we have \(f_{x,\lambda}(x) = \langle x_1, t \rangle\) for some \(x_1, t \in \mathbb{R}^n\). Let \(\alpha\) be the angle between \(x_1\) and \(x_2\), we then have

\[
\rho_X(X_{f,\lambda}) = \frac{1}{4} \frac{\|\text{sgn}(f_{x,\lambda}) - \text{sgn}((w_0, x))\|^2}{\rho} = \frac{1}{4} \frac{\|\text{sgn}(f_{x,\lambda}) - \text{sgn}(f_{x,\lambda})\|^2}{\rho} = \frac{\alpha}{\pi}. \tag{19}
\]

We will require the following result, a proof of which can also be found in [21], which makes use of the monotonicity of \(\hat{\eta}\).

Lemma 1. Let \(\hat{\eta}\) and \(\eta\) be as above, then

\[
\int_0^1 \hat{\eta}(t)(1 - t^2)^{\frac{n-3}{2}} dt \leq \int_0^1 \eta(t)(1 - t^2)^{\frac{n-3}{2}} dt. \tag{20}
\]

Proof. By definition, we have

\[
\eta = \frac{2|S^{n-2}|}{|S^{n-1}|} \int_0^1 \hat{\eta}(t)(1 - t^2)^{\frac{n-3}{2}} dt.
\]

Since \(|S^{n-1}| = 2|S^{n-2}| \int_0^1 (1 - t^2)^{\frac{n-3}{2}} dt\), it follows that

\[
\int_0^1 \hat{\eta}(t)(1 - t^2)^{\frac{n-3}{2}} dt = \int_0^1 \eta(t)(1 - t^2)^{\frac{n-3}{2}} dt = \frac{|S^{n-1}|}{2|S^{n-2}|} \eta.
\]

By symmetry of the inner product, this expression depends only on \(x_0\). It thus suffices to consider \(x\) of the form \(x = (0, \ldots, 0, \sqrt{1 - x_0^2}, x_0)\). We have

\[
\langle x, t \rangle = t_{n-1} \sqrt{1 - x_0^2} + t_n x_0.
\]

Let \(t_{n-1} = (\frac{1}{\sqrt{1-x_0^2}}, \ldots, \frac{1}{\sqrt{1-x_0^2}}) \in S^{n-2}\), \(e_{(n-1)} = (0, \ldots, 1) \in S^{n-2}\), then

\[
t = t_n e_n + \sqrt{1 - t_n^2} (t_{(n-1)}, 0)^T \quad \text{and} \quad x = x_n e_n + \sqrt{1 - x_n^2} (e_{(n-1)}, 0)^T.
\]

Thus it follows that

\[
\langle x, t \rangle = x_n t_n + \sqrt{1 - t_n^2} \sqrt{1 - x_n^2} (t_{(n-1)}, e_{(n-1)}).
\]
Using the property \( dS^{n-1} = (1 - t^2)^{\frac{n-3}{2}} dS^{n-2} dt \), we obtain
\[
\int \langle x, t \rangle [1 - 2\hat{\eta}(t)] dS^{n-1}(t) = \int_0^1 (1 - t^2)^{\frac{n-3}{2}} [1 - 2\hat{\eta}(t)] dt.
\]
By symmetry, we have
\[
\int \langle x, t \rangle dS^{n-1}(t) = -x_n |S^{n-2}| \int_0^1 (1 - t^2)^{\frac{n-3}{2}} t [1 - 2\hat{\eta}(t)] dt.
\]
It follows that, after normalization by dividing by \( |S^{n-1}| \),
\[
LK f_\rho(x) = 2 \frac{|S^{n-2}|}{|S^{n-1}|} x_n \int_0^1 (1 - t^2)^{\frac{n-3}{2}} t [1 - 2\hat{\eta}(t)] dt.
\]
From Lemma 1, we have
\[
\int_0^1 (1 - t^2)^{\frac{n-3}{2}} t [1 - 2\hat{\eta}(t)] dt \geq (1 - 2\eta) \int_0^1 (1 - t^2)^{\frac{n-3}{2}} t dt = \frac{1 - 2\eta}{n - 1}.
\]
Combining everything gives \( LK f_\rho(x) = Ax_n \), where \( A \geq \frac{2(1-2\eta)}{n-1} |S^{n-2}|/|S^{n-1}| \), as desired.

(b) The kernel \( K(x, t) = \langle x, t \rangle \) has exactly one eigenvalue \( \lambda_1 = \frac{1}{n} \) of multiplicity \( n \), with \( x_n \) being one eigenfunction. From the expression \( f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho \), we obtain \( f_\lambda \) \( \square \).

**Lemma 3.** Let \( w_1, w_2 \in \mathbb{R}^n \) with \( w_2 \neq 0 \) be such that \( \|w_1 - w_2\|/\|w_2\| \leq \sqrt{2} \). Let \( 0 \leq \alpha \leq \pi \) be the angle between \( w_1 \) and \( w_2 \). Then
\[
0 \leq \alpha \leq 2 \frac{\|w_1 - w_2\|}{\|w_2\|}.
\]
**Proof.** Let \( \|w_2\| = C \) and \( \|w_1 - w_2\| = D \). We have
\[
\|w_1 - w_2\|^2 \geq \|w_1\|^2 + \|w_2\|^2 - 2 \cos \alpha \|w_1\| \|w_2\| \Rightarrow \cos \alpha = \frac{\|w_1\|^2 + \|w_2\|^2 - \|w_1 - w_2\|^2}{2 \|w_1\| \|w_2\|} = \frac{\|w_1\|^2 + C^2 - D^2}{2 \|w_1\| C}.
\]

\[
\Rightarrow \sin^2 \alpha \leq \frac{D^2}{C^2} \leq \frac{D}{C}.
\]
It is straightforward to verify that on the interval \([0, \frac{\pi}{2}]\), we have \( \frac{1}{2} x \leq \sin x \leq x \). Thus with the assumption that \( 0 \leq \frac{D}{C} \leq \frac{\sqrt{2}}{2} \), then
\[
\sin \alpha \leq \frac{D}{C} \Rightarrow \alpha \leq 2D.
\]
as we claimed. \( \square \)

**Corollary 1.** Let \( z \) be randomly sampled according to \( \rho \) satisfying \( |y| \leq M \) almost surely. Suppose that \( f_{x_\lambda}(x) = \langle w_1, x \rangle \), \( f_{w_2}(x) = \langle w_2, x \rangle \) for some \( w_1, w_2 \in \mathbb{R}^n \). Then for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),
\[
\|w_1 - w_2\| \leq \frac{6K \lambda \log \frac{2}{\delta}}{\sqrt{m}}.
\]
**Proof.** This follows from Proposition 1 and the fact that \( \|w_1 - w_2\| = \|f_{x_\lambda} - f_{w_2}\| \), which is a consequence of the definition of the Weyl inner product. \( \square \)

**Proof of Theorem 1.** We have, from Lemma 2 and Corollary 1,
\[
\|w_2\| \geq \frac{2n(1 - 2\eta)}{1 + n\lambda} \frac{|S^{n-2}|}{|S^{n-1}|},
\]
and with probability at least \( 1 - \delta \),
\[
\|w_1 - w_2\| \leq \frac{6 \log \frac{2}{\delta}}{\sqrt{m} \lambda}.
\]
From Lemma 39 in [13], we have \( |S^{n-2}| > \frac{(n-1)^{1/2}}{2(n-1)!} \), hence for \( \lambda = \frac{1}{n} \), we have
\[
\|w_2\| > \frac{n(1 - 2\eta)}{e^{2/3} \sqrt{\pi (n-1)}} \quad \text{and} \quad \|w_1 - w_2\| \leq \frac{6 \log \frac{2}{\delta}}{\sqrt{m}}.
\]
Then \( \|w_1 - w_2\| \leq \sqrt{\frac{2}{3}} \|w_2\| \) if
\[
m \geq \frac{48e^{4/3} \pi (n-1)(\log \frac{2}{\delta})^2}{(1 - 2\eta)^2},
\]
which is satisfied if
\[
m \geq \frac{573 \lambda \log \frac{2}{\delta}}{2(n-1)(1 - 2\eta)^2}.
\]
Then 
\[
\alpha \leq \frac{2\|w_1 - w_2\|}{\|w_2\|} \leq \frac{12e^{2/3}\sqrt{\pi(n-1)}\log \frac{2}{\eta}}{\sqrt{m(1-2\eta)}}.
\]

It thus follows that 
\[
\rho_X(X_{f^*}) = \frac{\alpha}{\pi} \leq \frac{12e^{2/3}\sqrt{n-1}\log \frac{2}{\eta}}{\sqrt{m(1-2\eta)}} \leq \frac{14\sqrt{n}\log \frac{2}{\eta}}{\sqrt{m(1-2\eta)}},
\]
as we claimed. \( \square \)

6.3. The Average algorithm

In this section we prove Theorem 2 in the RKHS framework. We will need the following result.

**Theorem 3.** Let \( z \) be randomly drawn according to \( \rho \) satisfying \(|y| \leq M \) almost surely. Then for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),
\[
\frac{1}{m} \sum_{i=1}^{m} y_i K_{x_i} - L_K f_\rho \leq \frac{4M \log \frac{2}{\delta}}{\sqrt{m}}.
\]

**Proof of Theorem 2.** We apply Theorem 3 with \( K(x, t) = (x, t) \), noting that \( \text{sgn}(L_K f_\rho) = \text{sgn}(\langle w_0, x \rangle) \), and proceed as in the proof of Theorem 1. \( \square \)

To prove Theorem 3, we need the following law of large numbers for Hilbert space-valued random variables, which is a consequence of a general inequality due to Pinelis [15].

**Proposition 2.** (See [22].) Let \( H \) be a Hilbert space with norm \( \| \cdot \| \) and \( \xi \) be a random variable on \((Z, \rho)\) with values on \( H \). Assume that \( \|\xi\| \leq M < \infty \) almost surely for a fixed constant \( M > 0 \). Let \( \sigma^2(\xi) = \mathbb{E}(\|\xi\|^2) \). Let \( \{z_i\}_{i=1}^m \) be independently sampled according to \( \rho \). Then for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),
\[
\frac{1}{m} \sum_{i=1}^{m} \xi(z_i) - E\xi \leq \frac{2M \log \frac{2}{\delta}}{m} + \sqrt{\frac{2\sigma^2(\xi)\log \frac{2}{\delta}}{m}}.
\]

**Proof of Theorem 3.** Consider the random variable \( \xi : (Z, \rho) \to \mathcal{H}_K \) defined by \( \xi(x, y) = y K_x \). Then we have
\[
\frac{1}{m} \sum_{i=1}^{m} \xi(z_i) = \frac{1}{m} \sum_{i=1}^{m} y_i K_{x_i}
\]
and
\[
E(\xi) = \int X K_x \left( \int y d\rho(y|x) \right) d\rho_X(x) = L_K f_\rho.
\]

Also
\[
\|\xi(x, y)\|_K = |y| \sqrt{K(x, x)} \leq \kappa M \quad \text{and} \quad \sigma^2(\xi) \leq \kappa^2 M^2.
\]

By Proposition 2, for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),
\[
\frac{1}{m} \sum_{i=1}^{m} y_i K_{x_i} - L_K f_\rho \leq \frac{2M \log \frac{2}{\delta}}{m} + \sqrt{\frac{2\kappa^2 M^2 \log \frac{2}{\delta}}{m}},
\]
as desired. \( \square \)

7. Conclusion and future directions

We have presented the sample complexity for the half-space learning problem using the regularized least squares regression algorithm in RKHS, for the case the marginal distribution is uniform and under monotonic noise condition. We are aiming to improve our results and extend them to more general noise assumptions and marginal probability distributions. We hope that this will contribute to the theoretical understanding of the RKHS least squares method itself, particularly with regard to the general binary classification problem. Because of its generality and practical applicability, we also hope that it can also be applied to other problems of interest in computational learning theory.

References


