RISK BOUNDS FOR RANDOM REGRESSION GRAPHS

A. CAPONNETTO AND S. SMALE

Abstract. We consider the regression problem and describe an algorithm approximating the regression function by estimators piecewise constant on the elements of an adaptive partition. The partitions are iteratively constructed by suitable random merges and splits, using cuts of arbitrary geometry. We give a risk bound under the assumption that a "weak learning hypothesis" holds, and characterize this hypothesis in terms of a suitable RKHS.

1. Introduction

Many algorithms based on adaptive partitions have been proposed in the learning theory literature. Different algorithms adopt different strategies for the construction of the adaptive partitions. For example, probably the best known such an algorithm, CART [2], realizes a series of dyadic cuts by coordinate planes. In this case, the orientations of the cuts are chosen runtime and it is not set in advance how a particular cell will be split. On the contrary, other algorithms are based on a predetermined set of cuts [1], which constraints the geometry of the resulting cells, e.g. hypercubes. The aim of this paper is devising an algorithm which allows maximum freedom to the geometrical properties of the partitions.

The algorithm is based on an iterative procedure to construct a partition $\mathcal{P}$ of the input space X. It is defined in terms of a set $\mathcal{C}$ of allowed cuts of X, endowed with a probability measure π. At every iteration t a set of cuts are drawn i.i.d. from $\mathcal{C}$ according to π, and one of them is used to perform a split on the current partition. We have in mind (see Example 1) the case of cuts induced by arbitrary half-spaces in the Euclidean space E^d, randomly drawn according to an isotropic measure. In order to control the number of elements of the partition, the splits are followed by the merging of a class of (not necessarily connected) elements. The use of merges in the construction of adaptive partitions has been first proposed in the context of classification [14] and recently in the context of regression [12] [11]. But we are not aware of any previous thorough error analysis for adaptive algorithms involving both merges and splits by such general cuts.

A connection between this class of algorithms and boosting theory is also well established in the literature [13], and in fact our main error estimate (Theorem 1 in Subsection 2.3) relies on an assumption on the regression function f_ρ (Hypothesis 3), which can be described as a "weak learning hypothesis". It establishes a lower bound for the average squared covariance of f_ρ and the characteristic function of a randomly drawn half-space. Indeed, this quantity can be linked to the risk

Date: December 31, 2005.

2000 *Mathematics Subject Classification.* Primary 68T05, 68P30.

Key words and phrases. Regression graph, Risk bound, Reproducing kernel Hilbert space, Weak learning hypothesis.

Corresponding author: Andrea Caponnetto, +390103536609 (phone), +390103536699 (fax).

reduction, due to a random split, of the optimal piecewise constant estimators on the partitions (see Proposition A2). Therefore, Hypothesis 3 describes a guaranteed average risk reduction as effect of a random split.

The second main result of the paper (Theorem 2 in Section 3) shows that the functions belonging to the RKHS induced by the Mercer kernel

$$K(x,y) = 1 - 2C \left\| x - y \right\|,$$

satisfy Hypothesis 3.

From Theorem 1, Theorem 2, and the expression (4) for the constant C, we get the following Corollary.

Corollary 1. *Let the input space X be a ball of radius R in the d-dimensional Euclidean space, and consider the class of cuts $\mathcal{C}$ determined by general hyperplanes, endowed with the uniform measure π (see Example 1 in Subsection 2.1 for details). Assume that the regression function f_ρ fulfills*

$$\|f_\rho\|_{\mathcal{H}}^2 \leq V^2, \tag{1}$$

where $\mathcal{H}$ is the RKHS on X induced by the kernel

$$K(x,y) = 1 - \frac{2\Gamma(\frac{d-1}{2})}{\sqrt{\pi} R \Gamma(\frac{d}{2})} \left\| x - y \right\|.$$

Then, for every $m \geq (\frac{V}{4M})^2 + e$, with probability greater than $1 - 2\delta$, the following estimate for the expected risk of the estimator $\hat{f}$ returned by the Algorithm, holds

$$\left\| \hat{f} - f_\rho \right\|_\rho^2 = \mathcal{E}[\hat{f}] - \mathcal{E}[f_\rho] \leq 48 \max(M^2, V^2) \left(\frac{M^2}{V^2} \frac{\log(m/\delta)}{\sqrt{m}} \right)^{\frac{1}{3}}.$$

By Theorem B2 in Appendix B (see also Example $1'$ in Appendix B and the discussion in Example 1, Section 3), assumption (1) in the Corollary above can be replaced by the following condition

$$\frac{\pi}{2} R |S^{d-1}| \left\langle f_\rho^*, (-\triangle)^{\frac{d+1}{2}} f_\rho^* \right\rangle_{\mathcal{L}^2(E^d,\mu)} \leq V^2, \tag{2}$$

where $f_\rho^* \in \mathcal{L}^2(E^d)$ is an extension of f_ρ from X to the whole E^d, belonging to the smoothness space $H^{\frac{d+1}{2}}(E^d, \mu)$ defined in eq. (60) (here μ is the Lebesgue measure over the Euclidean space E^d). An asset of this result with respect to previous similar approaches (e.g. [12] and [11]) is that the space $H^{\frac{d+1}{2}}(E^d, \mu)$ is a *dense subspace* of $\mathcal{L}^2(E^d, \mu)$. However, the relation between the norm of the extension f_ρ^* in $H^{\frac{d+1}{2}}(E^d, \mu)$, which appears in eq. (2), and the original function f_ρ over X is not straightforward. Some estimates for the norm of f_ρ^* can be found for example in [3].

The paper is organized as follows. In Section 2, first, we define the setting of the learning problem, then we introduce the cuts $(\mathcal{C}, \pi)$ (Subsection 2.1), hence we describe the Algorithm (Subsection 2.2), and finally we give the main error estimate Theorem 1 (Subsection 2.3). In Section 3, Theorem 2 shows how the "weak learning hypothesis" (Hypothesis 3) can be characterized in terms of a suitable RKHS.

Throughout the paper Examples 1 and 2 illustrate the general results for two particularly interesting cases. In Example 1, the input space X is a ball in the Euclidean space E^d, and the cuts are induced by hyperplanes randomly drawn according to the uniform probability measure. Instead, Example 2 deals with the

hypersphere $X = S^{d-1}$, and considers cuts by isotropically distributed hyperplanes passing through the center of X.

Appendix A collects the preliminary results required for the proof of Theorem 1. Finally, Appendices B and C contain the results used in the analysis of Examples 1 and 2, respectively.

2. The learning problem

Let us first introduce the regression problem that we want to address. We assume the samples $z_i = (x_i, y_i) \in X \times Y$, $i = 1, \dots, m$, to be drawn i.i.d according to a probability measure ρ. Here $Y \subset [-M, M]$, for some known positive M.

The goal is determining, by means of the empirical samples $\mathbf{z} = (z_1, \dots, z_m)$, an estimator $\hat{f} := f_{\mathbf{z}} : X \to Y$ with low expected error

$$\mathcal{E}[\hat{f}] = \int_{X \times Y} (\hat{f}(x) - y)^2 d\rho(x, y).$$

In our context the empirical estimator $\hat{f}$ is defined by a partition of the input space X

$$\mathcal{P} = \{\ell_1, \dots, \ell_k\}.$$

The estimator $\hat{f}_{\mathcal{P}}$ induced by such a partition is given by

$$\hat{f}_{\mathcal{P}}(x) = \frac{1}{\hat{\rho}_X(\ell(x))} \int_{\ell(x) \times Y} y \; d\hat{\rho},$$

where $\ell(x)$ is uniquely defined by the condition $x \in \ell(x) \in \mathcal{P}$, and the empirical distribution $\hat{\rho}$ is defined in terms of the empirical samples by

$$\hat{\rho} = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i} \delta_{y_i},$$

and $\hat{\rho}_X$ is the corresponding marginal distribution over X.

In Subsection 2.2 we will describe the algorithmic construction of $\hat{f}$. Hence, in Subsection 2.3 we will state the main result of this section: an upper bound on the expected error $\mathcal{E}[\hat{f}]$. But before we have to introduce the main concept involved in the construction of the partitions of X. That is, the set of allowed splits of X. This is the topic of the next subsection.

2.1. Splits.

As it will be thoroughly explained in the next subsection, the main elementary procedure in the construction of the partitions, is the splitting of subsets of X by cuts randomly drawn from a class. An allowed split of $\ell \subset X$ generates the two parts $\ell \cap c$ and $\ell \cap \bar{c}$, where the subset of X, c, is drawn from the class of subsets $\mathcal{C}$ (here $\bar{c}$ is the complement of c in X). Since the Algorithm will implement random splits we must also endow the class $\mathcal{C}$ with a probability measure π.

More formally, let $\mathcal{C}$ be a set of subsets of X and $(\mathcal{C}, \pi, \mathcal{F})$ a probability space on it. It is useful assuming that the following hypothesis holds

Hypothesis 1. $\forall x \in X$, $\mathcal{C}_x := \{c \in \mathcal{C} |\ x \in c\} \in \mathcal{F}$.

For the following developments it is also useful showing how the probability space $(\mathcal{C}, \pi)$ induces a natural metric structure over the input space X.

In particular we study the properties of the kernel $D : X \times X \to \mathbb{R}$ defined by

Definition 1.

$$D(x,y) := \pi(\mathcal{C}_x \cap \bar{\mathcal{C}}_y) + \pi(\mathcal{C}_y \cap \bar{\mathcal{C}}_x),$$

where, by $\bar{S}$ we denoted the complement of S in $\mathcal{C}$. Clearly $D(x,y)$ represents the probability that a random cut separates the point x from y. Moreover, as shown by the following proposition, D is a distance function over X.

Proposition 1. *For all $x, y \in X$, $D(x,y)$ fulfills the following properties*

(1) $D(x,y) \geq 0$,
(2) $D(x,y) = D(y,x)$,
(3) $D(x,y) \leq D(x,z) + D(z,y)\ \forall z \in X$.

Proof. Non-negativity and symmetry are obvious by the definition of D. Triangle inequality can be derived noticing that

$$\mathcal{C}_x \cap \bar{\mathcal{C}}_y = (\mathcal{C}_x \cap \bar{\mathcal{C}}_y) \cap (\mathcal{C}_z \cup \bar{\mathcal{C}}_z) \subset (\mathcal{C}_z \cap \bar{\mathcal{C}}_y) \cup (\mathcal{C}_x \cap \bar{\mathcal{C}}_z).$$

Hence by union bound

$$\pi(\mathcal{C}_x \cap \bar{\mathcal{C}}_y) \leq \pi(\mathcal{C}_x \cap \bar{\mathcal{C}}_z) + \pi(\mathcal{C}_z \cap \bar{\mathcal{C}}_y),$$

which added to the analogous inequality obtained by switching x and y gives the desired result. □

In a strict sense, Proposition 1 shows that (X, D) is a pseudo-metric space, since $D(x,y) = 0$ does not imply $x = y$. But also in the general case, it is possible to recover the familiar metric structure working with suitable equivalence classes of points in X. However, in all the examples that we will consider, D is already a metric on X. Hence, hereafter we assume that indeed (X, D) is a metric space.

For the further developments we also assume that the the metric space (X, D) fulfills the additional technical hypothesis

Hypothesis 2.

(1) (X, D) is separable,
(2) the elements of $\mathcal{C}$ are Borel sets of (X, D).

Now, let us illustrate two instances of the structure $(\mathcal{C}, \pi)$.

Example 1 Let X be the closed ball with center o and radius R in the d-dimensional Euclidean space. Define

$$\mathcal{C} := \{c(\omega, p) |\ (\omega, p) \in S^{d-1} \times (-R, R)\},$$

where

$$c(\omega, p) := \{x \in X |\ \omega \cdot (o - x) > p\}.$$

Endow $\mathcal{C}$ with the σ-field $\mathcal{F}$ and the measure π induced by the natural product measure over $S^{d-1} \times (-R, R)$, that is $d\pi = dpd\omega/(2R|S^{d-1}|)$. Clearly Hypothesis 1 is fulfilled.

The kernel D can be obtained by direct computation as follows

$$D(x,y) = 2\pi(\mathcal{C}_x \cap \bar{\mathcal{C}}_y) \tag{3}$$

$$= \frac{1}{R|S^{d-1}|} \int_{S^{d-1} \times (-R,R)} \chi_{\{\omega \cdot (o-x) > p\}} \chi_{\{\omega \cdot (o-y) \leq p\}} = \frac{1}{R|S^{d-1}|} \int_{S^{d-1}} |\ \omega \cdot (x - y)| d\omega$$

$$= \frac{|S^{d-2}|}{R|S^{d-1}|} \|x - y\| \int_0^\pi (\cos\theta)^{d-2} \sin\theta d\theta = C(d, R) \|x - y\|,$$

where

$$C(d,R)=\frac{\Gamma(\frac{d-1}{2})}{\sqrt{\pi}R\Gamma(\frac{d}{2})}=\sqrt{\frac{2}{\pi R^2 d}}+O(d^{-\frac{3}{2}}). \tag{4}$$

Clearly, since $D(x,y)$ is proportional to the Euclidean distance between x and y, Hypothesis 2 is fulfilled.

Example 2 Let X be the $(d-1)$-dimensional sphere of unit radius S^{d-1}. Define

$$\mathcal{C}:=\{c(\omega)|\ \omega\in X\},$$

where

$$c(\omega):=\{x\in X|\ d(\omega,x)<\pi/2\},$$

with $d(\cdot,\cdot)$ the geodesic distance over S^{d-1}.

Endow $\mathcal{C}$ with the σ-field $\mathcal{F}$ and the measure π induced by the natural normalized measure over S^{d-1}. Clearly Hypothesis 1 is fulfilled.

In order to compute $D(x,y)$, identify X with the sphere of unit radius and center o in E^d. In general, o, x and y single out a two-dimensional linear manifold. Name $\ell=\ell(x,y)$ the circle obtained intersecting this plane with the sphere. Now, consider a general $\omega\in X$ and the corresponding hemisphere $c(\omega)$. Its border $\partial c(\omega)$ is given by the points z on the sphere such that $\omega\cdot z=0$. Reasoning on the three-dimensional linear manifold singled out by o, x, y and ω it is easy to verify that, generally, $\partial c(\omega)\cap\ell$ is a set of two antipodal points $\{\pm p(\omega,\ell)\}$. Now, observe that $c(\omega)\in(\mathcal{C}_x\cap\bar{\mathcal{C}}_y)\cup(\mathcal{C}_y\cap\bar{\mathcal{C}}_x)$ if and only if x and y belong to the two different hemispheres of border $\partial c(\omega)$. In this case any continuous line connecting x and y (and hence also, $\ell_<$, the smallest arc on ℓ with these end points) must intersect $\partial c(\omega)$. Hence

$$D(x,y)=\pi((\mathcal{C}_x\cap\bar{\mathcal{C}}_y)\cup(\mathcal{C}_y\cap\bar{\mathcal{C}}_x))=\pi(c(\omega)|\pm p(\omega,\ell)\in\ell_<).$$

Finally we observe that the random variable $p(\omega,\ell)$ is uniformly distributed over the circle ℓ. In fact, by symmetry, its measure must be invariant with respect to any subgroup of rotations of the circle ℓ. Only the Lebesgue measure over S^1, μ, has such property (modulo normalization). Normalizing (for antipodal points $D(x,y)=1$) we obtain

$$D(x,y)=2\frac{\mu(\ell_<)}{2\pi}=\frac{1}{\pi}d(x,y).$$

Clearly, since $D(x,y)$ is proportional to the geodesic distance between x and y, Hypothesis 2 is fulfilled.

2.2. **Description of the algorithm.** The algorithm that we consider works iteratively. The partition at time t is obtained by a suitable transformation $\hat{A}_t:=A_{\mathbf{z},t}$ of the partition at time $t-1$, that is

$$\mathcal{P}_t=\hat{A}_t\mathcal{P}_{t-1},$$

with the initial condition $\mathcal{P}_0=\{X\}$. At time $T=T(m)$ (defined in eq. (8)), the algorithm stops and outputs the final estimator $\hat{f}=\hat{f}_{\mathcal{P}_T}$.

The transformation $\hat{A}_t$ is designed with the aim of reducing the empirical error associated with the new partition

$$\hat{\mathcal{E}}[\mathcal{P}]:=\hat{\mathcal{E}}[\hat{f}_{\mathcal{P}}]:=\frac{1}{m}\sum_{i=1}^{m}(\hat{f}_{\mathcal{P}}(x_i)-y_i)^2.$$

The *elementary transformations* we start from, are simple merge-split steps which we denote by $\hat{A}^B_{h,c}$. Here, $c \in \mathcal{C}$, B is a positive integer, and h is an integer in $\{1, \ldots, 2B\}$. The parameter B induces a new partition $\mathcal{B} = \{b_h\}_{h=1,\ldots,2B}$, with elements defined by

$$b_h = \{x \in X |\ M(h-1) < B\hat{f}_{\mathcal{P}}(x) + MB \leq Mh\},\ h = 2, \ldots, 2B.$$

Since $\hat{f}_{\mathcal{P}}$ is piecewise constant on the elements of $\mathcal{P}$, $\mathcal{B}$ is clearly coarser that $\mathcal{P}$.

`Action of` $\hat{A}^B_{h,c}$ `on` $\mathcal{P}$`.`

· merge: replace $\{\ell \in \mathcal{P} |\ \ell \subset b_h\}$ with $\{b_h\}$.
· split: replace $\{b_h\}$ with $\{b_h \cap c, b_h \cap \bar{c}\}$.

The transformation $\hat{A}_t$ is realized choosing, among a suitably constructed set of elementary transformations, the one leading to the largest empirical error reduction. The elementary transformations at time t are induced by the set $\Omega_t \in \mathcal{C}^{s_t}$ of elements in $\mathcal{C}$ drawn i.i.d. according to the probability distribution π (in practice this procedure is realized by a suitable parametrization of $\mathcal{C}$, e.g. $(\omega, p) \in S^{d-1} \times (-R, R)$ in Example 1), and s_t is the increasing function,

$$s_t = \left\lceil 4\psi_t^{-\frac{1}{2}} \log \frac{6\psi_t^{-\frac{5}{2}}}{5\delta} \right\rceil,\ \text{with } 0 < \delta \leq 1, \tag{5}$$

where $\lceil x \rceil$ is the smallest integer greater or equal to x, and

$$\psi_t = \frac{1}{2t+5}. \tag{6}$$

The allowed elementary transformations considered at iteration t are the $\hat{A}^{B_t}_{h,c}$, where $c \in \Omega_t$ and $h \in \{1, \ldots, 2B_t\}$ with,

$$B_t = \left\lceil \psi_t^{-1/2} \right\rceil. \tag{7}$$

At every iteration t between 1 and T, the new partition $\mathcal{P}_t$ is chosen, among the candidates $\hat{A}^{B_t}_{h,c}\mathcal{P}_{t-1}$, in order to minimize the empirical error $\hat{\mathcal{E}}[\mathcal{P}_t]$. The Algorithm is illustrated by the pseudo-code below.

`Algorithm.`

· $\mathcal{P}_0 = \{X\}$.
· for t from 1 to $T = T(m)$.
· $(h^*, c^*) = \mathrm{argmin}_{(h,c) \in \{1,\ldots,2B_t\} \times \Omega_t} \hat{\mathcal{E}}[\hat{A}^{B_t}_{h,c}\mathcal{P}_{t-1}]$.
· $\mathcal{P}_t = \hat{A}^{B_t}_{h^*,c^*}\mathcal{P}_{t-1}$.
· output $\hat{f} = \hat{f}_{\mathcal{P}_T}$.

The stopping time $T = T(m)$ is defined by

$$T(m) = \left\lfloor \frac{1}{32} \left(\frac{V^4}{M^4} \frac{m}{\log^2(m/\delta)} \right)^{\frac{1}{3}} - \frac{5}{2} \right\rfloor, \tag{8}$$

where $\lfloor x \rfloor$ is the largest integer smaller or equal to x.

It is worth noticing that the actual value of V appears in the definition of the algorithm, only through the expression of the stopping time $T(m)$. Even if an

estimate of V is usually not available, in practice the stopping time can be chosen using a cross-validation technique, without any relevant reduction of performance [7] [4].

2.3. **Error analysis.** The error analysis that we propose is based on the following assumption on the distribution ρ. We will discuss a possible characterization of this hypothesis in Section 3.

Hypothesis 3. We assume, with reference to the framework of the previous subsections, that there exists a positive constant V, such that for every probability measure ν over (X, D), it holds

$$\mathrm{var}^2_\nu(f_\rho) \leq 4V^2 \int_{\mathcal{C}} \mathrm{cov}^2_\nu(f_\rho, \chi_c) d\pi(c),$$

where $f_\rho(x) = \mathbb{E}_{\rho(y|x)} y$.

The main error estimate of this section is given in the Theorem below.

Theorem 1. *Let us assume that the probability measure ρ fulfills Hypothesis 3. Then, for every $m > (\frac{V}{4M})^2 + e$, with probability greater than $1 - 2\delta$ the expected risk of the estimator $\hat{f}$ returned by the Algorithm, fulfills*

$$\left\| \hat{f} - f_\rho \right\|^2_\rho = \mathcal{E}[\hat{f}] - \mathcal{E}[f_\rho] \leq 48 \max(M^2, V^2) \left(\frac{M^2}{V^2} \frac{\log(m/\delta)}{\sqrt{m}} \right)^{\frac{1}{3}}.$$

We now prove Theorem 1, we use Propositions 2, 3, 4 and 5, whose proofs are given in Appendix A.

Proof. First let us establish a concentration result of empirical risks to expected risks. We will reason conditionally with respect to the sequence $\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_T)$.

We need some more notation. Let Θ_0 be the trivial partition $\{\mathcal{P}_0\}$, and Θ_t (for $t \geq 1$) the class of partitions obtained from the elements of Θ_{t-1}, by arbitrary merges of some subsets of leaves, followed by a single split $\{b\} \to \{b \cap c, b \cap \bar{c}\}$, with $c \in \Omega_t$. Moreover, let Θ'_t (for $t \geq 1$) be the class of partitions obtained from the elements $\mathcal{P} \in \Theta_{t-1}$, by merges of arbitrary subsets of leaves $\mathcal{P} \to \mathcal{B}$, and subsequent splits $\{b_i\} \to \{b_i \cap c_i, b_i \cap \bar{c}_i\}$ (again with cuts $c_i \in \Omega_t$) of an arbitrary subset of elements $b_i \in \mathcal{B}$.

Finally, for every partition $\mathcal{P}$ of X, let $\mathcal{F}[\mathcal{P}]$ be the class of functions from X to $[-M, M]$ piecewise constant on the elements of $\mathcal{P}$. We have the following concentration result.

Proposition 2. *Let $\tilde{\Theta}_T := \bigcup_{t=1}^T \Theta'_t$. For every $m \in \mathbb{N}$ and $0 < \delta \leq 1$,*

$$\Pr_{\mathbf{z} \sim \rho^m} \left(\sup_{f \in \mathcal{F}[\tilde{\Theta}_T]} |\hat{\mathcal{E}}[f] - \mathcal{E}[f]| > q(m, \delta) \right) \leq \delta, \tag{9}$$

where

$$q(m, \delta) = 4M^2 \sqrt{\frac{2T \log(8mTs_T) + \log(2/\delta)}{m}} + \frac{4M^2}{m}. \tag{10}$$

Since $\mathbf{z}$ and θ are independent, from Eq. (9), and the relation $\Pr_{\mathbf{z},\Omega} A = \mathbb{E}_\Omega \Pr_{\mathbf{z} \sim \rho^m} A$, it follows straightforwardly,

$$\Pr_{\mathbf{z},\Omega} \left(\sup_{f \in \mathcal{F}[\tilde{\Theta}_T]} |\hat{\mathcal{E}}[f] - \mathcal{E}[f]| > q(m, \delta) \right) \leq \delta.$$

Now, it is clear that $\mathcal{P}_t \in \Theta_t$ for every $t = 0, \ldots, T$, hence $f_{\mathcal{P}_t}$ and $\hat{f}_{\mathcal{P}_t}$ belong to $\mathcal{F}[\tilde{\Theta}_T]$. Hence, from Proposition 2, we get

$$\Pr_{\mathbf{z},\Omega}\left(\mathcal{E}[\hat{f}] > \hat{\mathcal{E}}[\hat{f}] + q(m,\delta)\right) \leq \delta. \tag{11}$$

The following step is to control the decrease of empirical error $\hat{\mathcal{E}}[\mathcal{P}_t]$ at every t. This is accomplished by the following Proposition.

Proposition 3. *Let us use the notation introduced in the previous subsections, and assume that the samples $\mathbf{z} \in (X \times Y)^m$ are drawn i.i.d. according to a distribution ρ fulfilling Hypothesis 3. Moreover let $\hat{\xi}_t := \hat{\mathcal{E}}[\mathcal{P}_t] - \mathcal{E}[f_\rho]$, and $w_t := (\pi^2 t^2 B_t/3\delta)^{1/s_t} - 1$. Then, with probability greater than $1 - 2\delta$, for every $t = 1, \ldots, T$, it holds*

$$\begin{aligned} &\hat{\mathcal{E}}[\mathcal{P}_{t-1}] - \min_{(h,c)\in\{1,\ldots,2B_t\}\times\Omega_t} \hat{\mathcal{E}}[\hat{A}^{B_t}_{h,c}\mathcal{P}_{t-1}] \\ &\geq 2C_t^{(2)}\hat{\xi}_{t-1}^2 - \frac{1}{2}C_t^{(1)}\hat{\xi}_{t-1} - \frac{1}{2}C_t^{(0)}, \end{aligned} \tag{12}$$

where,

$$\begin{aligned} C_t^{(0)} &= B_t^{-1}\left((1+w_t)M^2 B_t^{-2} + (2 + w_t + 3V^{-2}q(m,\delta))q(m,\delta)\right), \\ C_t^{(1)} &= B_t^{-1}\left(w_t + 2q(m,\delta)V^{-2}\right), \\ C_t^{(2)} &= (4\max(V^2, M^2)B_t)^{-1}. \end{aligned}$$

By the definition of the Algorithm, we have that $\hat{\mathcal{E}}[\mathcal{P}_t] = \min_{(h,c)\in\{1,\ldots,2B_t\}\times\Omega_t} \hat{\mathcal{E}}[\hat{A}^{B_t}_{h,c}\mathcal{P}_{t-1}]$. Therefore inequality (12) becomes

$$\hat{\xi}_t \leq \hat{\xi}_{t-1} - 2C_t^{(2)}\hat{\xi}_{t-1}^2 + \frac{1}{2}C_t^{(1)}\hat{\xi}_{t-1} + \frac{1}{2}C_t^{(0)}.$$

The inequality above gives a lower bound for the reduction of $\hat{\xi}_t$ at every iteration t. In order to pass from this incremental result to an upper bound on $\hat{\xi}_t$ itself, we will apply Proposition 4 below. But we first have to transform the inequality above into the form of (16). This is achieved by multiplying both sides by $C_t^{(2)}$, using the fact that B_t increases with t, and using the identifications

$$e_{t-1} := C_t^{(2)}\hat{\xi}_{t-1}, \tag{13}$$

$$\psi_1(2(t+2)) := C_t^{(1)}, \tag{14}$$

$$\psi_2(2(t+2))^2 := C_t^{(2)}C_t^{(0)}. \tag{15}$$

Proposition 4. *Let $e_0 \leq e/(1+3e)$ for some $e \geq 0$. Assume that for every integer $t \geq 1$,*

$$e_t \leq e_{t-1} - 2e_{t-1}^2 + \frac{1}{2}e_{t-1}\ \psi_1(2(t+2)) + \frac{1}{2}\psi_2(2(t+2))^2, \tag{16}$$

with,

$$0 \leq \psi_i(t) \leq \psi(t) := \frac{e}{1+et},\ i = 1, 2. \tag{17}$$

Then, for all $t \geq 1$,

$$e_t \leq \psi(t+3).$$

In order to apply Proposition 4, we must verify the initial condition $e_0 \leq e/(1+3e)$ and the validity of the bounds (17).

The initial condition can be easily verified observing that $e_{t-1} = C_t^{(2)}\hat{\xi}_{t-1} \leq \hat{\mathcal{E}}[\mathcal{P}_{t-1}]/(4M^2) \leq 1/4$, which is consistent with the choice $e = 1$.

Indeed, Proposition 5 below shows that conditions (17) are verified.

Proposition 5. *For every $t \geq 0$, the expressions (5) and (7) for s_t and B_t, and the constraint ($q(m,\delta)$ is defined in (10)),*

$$q(m,\delta) \leq \frac{1}{4}V^2\psi_T, \tag{18}$$

imply,

$$C_t^{(1)} \leq \psi_t, \tag{19}$$

$$C_t^{(2)}C_t^{(0)} \leq \psi_t^2. \tag{20}$$

Moreover for $m \geq (\frac{V}{4M})^2 + e$, the constraint (18) is enforced by the choice of $T(m)$ given in (8).

Hence, from Proposition 4 it follows

$$e_{t-1} \leq \psi(t+2) = \frac{1}{t+3},\ \forall t = 1,\dots,T, \tag{21}$$

and recalling eq. (13), eq. (11) and eq. (18), we get

$$\begin{aligned}
\mathcal{E}[\hat{f}] - \mathcal{E}[f_\rho] &\leq \hat{\xi}_T + q(m,\delta) \leq \frac{4\max(V^2,M^2)B_T}{T+4} + \frac{V^2}{4}\psi_T \\
&\leq \max(V^2,M^2)\left(8\psi_{T+1}(\psi_{T+1}^{-\frac{1}{2}}+1) + \frac{3}{10}\psi_{T+1}\right) \\
&\leq 12\max(V^2,M^2)\psi_{T+1}^{\frac{1}{2}} \\
&\leq 48\max(V^2,M^2)\left(\frac{M^2}{V^2}\frac{\log(m/\delta)}{\sqrt{m}}\right)^{\frac{1}{3}},
\end{aligned}$$

which concludes the proof. □

3. Covariance estimate on RKHS

In this section we show a connection between Hypothesis 3 and a suitable RKHS. We begin by showing how to construct a RKHS of continuous functions over the metric space (X,D), from the probability space $(\mathcal{C},\pi)$ fulfilling Hypotheses 1 and 2.

Proposition 6. *The kernel K defined by*

$$K(x,y) = 1 - 2D(x,y),$$

is a Mercer kernel over the metric space (X,D).

Proof. Symmetry and continuity relative to the topology induced by D are obvious. Positive-definiteness follows by the representation

$$K(x,y) = \int_{\mathcal{C}} \phi_x\phi_y d\pi, \tag{22}$$

where

$$\phi_x(c) := \chi_c(x) - \chi_{\bar{c}}(x) \ \forall c \in \mathcal{C}, \tag{23}$$

and χ_c is the characteristic function of the set c. The representation above can be verified noticing that by the definition of ϕ_x

$$\begin{aligned} \int_{\mathcal{C}} \phi_x \phi_y d\pi &= \pi((\mathcal{C}_x \cap \mathcal{C}_y) \cup (\bar{\mathcal{C}_x} \cap \bar{\mathcal{C}_y})) - \pi((\mathcal{C}_x \cap \bar{\mathcal{C}_y}) \cup (\bar{\mathcal{C}_x} \cap \mathcal{C}_y)) \\ &= \pi(\mathcal{C}) - 2\pi((\mathcal{C}_x \cap \bar{\mathcal{C}_y}) \cup (\bar{\mathcal{C}_x} \cap \mathcal{C}_y)) = 1 - 2D(x,y). \end{aligned}$$

□

The uniformly bounded Mercer kernel K induces a RKHS of continuous functions over the separable (by Hypothesis 2) metric space (X, D), which we name $\mathcal{H}$ ([5], [6]).

Let us now consider an arbitrary non-degenerate probability measure ν on (X, D). We will consider the bounded self-adjoint linear operator
$L_K : L^2(X,\nu) \to L^2(X,\nu)$ defined by

$$(L_K f)(x) = \int_X K(x,y) f(y) d\nu(y).$$

It is a well-known fact that $L_K^{-1/2}$ defines an isometry from $\mathcal{H}$ to $L^2(X,\nu)$ ([5], [6]), that is

$$\|f\|_{\mathcal{H}} = \left\| L_K^{-1/2} f \right\|_\nu, \ \forall f \in \mathcal{H}. \tag{24}$$

This property is useful in the proof of the following proposition which establishes a lower bound for the average squared covariance of a function f in $\mathcal{H}$ and the characteristic function of a random element in $\mathcal{C}$.

Theorem 2. *For every $f \in \mathcal{H}$, defining $V^2(f) := \|f\|_{\mathcal{H}}^2$, the estimate below holds*

$$\mathrm{var}_\nu^2(f) \leq 4V^2(f) \int_{\mathcal{C}} \mathrm{cov}_\nu^2(f, \chi_c) d\pi(c), \tag{25}$$

where the variance and covariance are relative to an arbitrary probability measure ν over (X, D).

Proof. Let P_0 be the orthogonal projector in $L^2(X,\nu)$ over the linear subspace of zero mean functions, that is

$$P_0 f = f - \mathbf{1} \langle f, \mathbf{1} \rangle_\nu .$$

Clearly, due to the properties of orthogonal projectors, we can write

$$\mathrm{cov}(f,g) = \langle P_0 f, P_0 g \rangle_\nu = \left\langle f, P_0^2 g \right\rangle_\nu = \langle f, P_0 g \rangle_\nu . \tag{26}$$

Hence using Cauchy-Schwartz inequality and equation (24) we obtain

$$\begin{aligned} \mathrm{var}^2(f) &= \langle f, P_0 f \rangle_\nu^2 = \left\langle L_K^{1/2} L_K^{-1/2} f, P_0 f \right\rangle_\nu^2 \\ &= \left\langle L_K^{-1/2} f, L_K^{1/2} P_0 f \right\rangle_\nu^2 \leq \|f\|_{\mathcal{H}}^2 \langle L_K P_0 f, P_0 f \rangle_\nu . \end{aligned} \tag{27}$$

Now we use definition (1) and equation (26) to estimate the term $\langle L_K P_0 f, P_0 f\rangle_\nu$, we obtain

$$\langle L_K P_0 f, P_0 f\rangle_\nu = \langle \mathbf{1}, P_0 f\rangle_\nu^2 - 2\int_{X\times X} (P_0 f)(x) D(x,y)(P_0 f)(y) d\nu^2(x,y)$$

$$= -2\int_{X\times X} (P_0 f)(x)\left(\int_{\mathcal{C}} \chi_c(x)\chi_{\bar{c}}(y) + \chi_{\bar{c}}(x)\chi_c(y)\ d\pi(c)\right)(P_0 f)(y) d\nu^2(x,y)$$

$$= -4\int_{\mathcal{C}} \langle P_0 f, \chi_c\rangle_\nu \langle P_0 f, \chi_{\bar{c}}\rangle_\nu\ d\pi(c) = -4\int_{\mathcal{C}} \langle P_0 f, \chi_c\rangle_\nu \langle P_0 f, \mathbf{1}-\chi_c\rangle_\nu\ d\pi(c)$$

$$= 4\int_{\mathcal{C}} \langle P_0 f, \chi_c\rangle_\nu^2\ d\pi(c) = 4\int_{\mathcal{C}} \mathrm{cov}^2(f,\chi_c)\ d\pi(c), \tag{28}$$

where all scalar products are well defined since, by Hypothesis 2, the characteristic functions χ_c are integrable.

The proposition follows from (27) and (28). □

We conclude this part analyzing the nature of the RKHS norm $\|\cdot\|_{\mathcal{H}}$ for the examples we described in the previous section.

Example 1 From Theorem 2 it is clear that Hypothesis 3 holds for $f_\rho \in \mathcal{H}$, with

$$V^2 = \|f_\rho\|_{\mathcal{H}}^2,$$

where $\mathcal{H}$ is the RKHS induced by the kernel $K(x,y) = 1 - 2C(d,R)\|x-y\|$ over the ball of radius R in the d-dimensional Euclidean space.

It is also possible to obtain a more explicit estimate for V^2 in terms of the smoothness properties of f_ρ. In fact (see Appendix B, Thm. B2), for odd d, V^2 can be characterized in terms of the minimal norm in $H^{\frac{d+1}{2}}(E^d)$ (see definition (60)) attained by extensions of f_ρ to the whole E^d. Indeed, it is possible to prove that, if f_ρ can be extended to a function in $H^{\frac{d+1}{2}}(E^d)$, then the minimum norm extension f_ρ^* exists and is unique. Now, since ν has support over the ball of radius R, $\mathrm{cov}_\nu^2(f_\rho^*, \chi_c) = 0$ for $c = c(\omega, p)$ with $|p| > R$. Therefore, from the estimate (62), normalizing, we obtain that Hypothesis 3 holds with

$$V^2 = \frac{\pi}{2} R |S^{d-1}| \left\langle f_\rho^*, (-\triangle)^{\frac{d+1}{2}} f_\rho^* \right\rangle_{\mathcal{L}^2(E^d,\mu)}. \tag{29}$$

Example 2 The kernel K over the sphere S^{d-1} has the form $K(x,y) = 1 - \frac{2}{\pi} d(x,y)$. It can be shown that the corresponding RKHS $\mathcal{H}$ is given, for even $d \geq 4$, by the antisymmetric functions in the Sobolev space $H^{d/2}(S^{d-1}) \subset \mathcal{L}^2(S^{d-1})$ (see Appendix C, Thm. C1). For $f_\rho \in \mathcal{H}$, by Thm. 2 and Thm. C1, we have that Hypothesis 3 is fulfilled with

$$V^2 = C\left\langle f_\rho, P_{d/2}(-\triangle) f_\rho\right\rangle_{\mathcal{L}^2(S^{d-1})},$$

where $P_{d/2}$ and C defined in (63) and (64).

Appendix A. Proofs of Propositions 2, 3, 4, and 5

Proof of Proposition 2. For every $f : X \to [-M, M]$ let $\dot{f}$ be the function obtained rounding f to the values

$$v_i = \frac{M}{2m}(2i - 2m - 1),\ i = 1, \ldots, 2m$$

clearly, since $2|f(x) - \dot{f}(x)| \leq M/m$, and,

$$|(\dot{f}(x) - f_\rho(x))^2 - (f(x) - f_\rho(x))^2| \leq |(f(x) - \dot{f}(x))(f(x) + \dot{f}(x) - 2f_\rho(x))| \leq 2M^2/m, \tag{30}$$

one has,

$$\begin{aligned} |\hat{\mathcal{E}}[f] - \mathcal{E}[f]| &\leq |\hat{\mathcal{E}}[f] - \hat{\mathcal{E}}[\dot{f}]| + |\hat{\mathcal{E}}[\dot{f}] - \mathcal{E}[\dot{f}]| + |\mathcal{E}[\dot{f}] - \mathcal{E}[f]| \\ &\leq 4M^2/m + |\hat{\mathcal{E}}[\dot{f}] - \mathcal{E}[\dot{f}]|. \end{aligned} \tag{31}$$

Denote by $\dot{\mathcal{F}}(\mathcal{P})$ the functions piecewise constant on the elements of $\mathcal{P}$ and taking values over $\{v_i\}_{i=1}^{2m}$. Since $\#\dot{\mathcal{F}}(\mathcal{P}) = (2m)^{\#\mathcal{P}}$ and $\#\mathcal{P} \leq 2T$ for every $\mathcal{P} \in \tilde{\Theta}_T$, from Proposition A1 it follows

$$\#\dot{\mathcal{F}}(\tilde{\Theta}_T) \leq (8mTs_T)^{2T}. \tag{32}$$

Finally, applying Eqn. (31) and Hoeffding inequality [10],

$$\begin{aligned} & \Pr_{\mathbf{z}\sim\rho^m}\left(\sup_{f\in\mathcal{F}[\tilde{\Theta}_T]} |\hat{\mathcal{E}}[f] - \mathcal{E}[f]| > 4M^2\sqrt{\frac{\log(2/\delta')}{m}} + 4M^2/m\right) \\ \leq & \Pr_{\mathbf{z}\sim\rho^m}\left(\sup_{f\in\dot{\mathcal{F}}[\tilde{\Theta}_T]} |\hat{\mathcal{E}}[f] - \mathcal{E}[f]| > 4M^2\sqrt{\frac{\log(2/\delta')}{m}}\right) \\ \leq & \sum_{f\in\dot{\mathcal{F}}[\tilde{\Theta}_T]} \Pr_{\mathbf{z}\sim\rho^m}\left(|\hat{\mathcal{E}}[f] - \mathcal{E}[f]| > 4M^2\sqrt{\frac{\log(2/\delta')}{m}}\right) \\ \leq & \#\dot{\mathcal{F}}(\tilde{\Theta}_T)\delta', \end{aligned}$$

which, letting $\delta = \delta'(8mTs_T)^{2T}$, by Eqn. (32) proves the Proposition. □

Proposition A1. *It holds*

$$\#\bigcup_{t=1}^{T}\Theta'_t \leq (4Ts_T)^{2T}.$$

Proof. We represent the partitions by directed acyclic graphs with root node. The nodes with outgoing edges are annotated with a set $c \in \mathcal{C}$. Every annotated graph has two outgoing edges labeled with the two truth values. Each $x \in X$ is mapped to a leaf of the graph, by beginning at the root, and following, at every annotated node, the outgoing edge labeled by the truth value of the binary predicate $\{x \in c\}$. It is clear that this mapping induces a partition of X. We want to show that any partition in Θ_t can be represented by a graph with exactly t annotated nodes (with annotations $c_i \in \Omega_i$, $i = 1, \ldots, t$) and $t + 1$ leaves.

Let us reason by induction. $\mathcal{P}_0$ is represented by the root node alone, which proves the thesis for Θ_0. Now, assume the thesis is true for Θ_{t-1}. Beginning with a graph representing a partition in Θ_{t-1}, the merges are achieved by rearranging the incoming edges of the leaves, which keeps unchanged the number of nodes and leaves. The split is achieved by annotating one of the leaves with $c_t \in \Omega_t$ and adding two new leaves. Hence the new graph, representing an arbitrary partition in Θ_t, has exactly t annotated nodes and $t + 1$ leaves, as claimed.

Now, from the definition of Θ'_t, it is clear that any partition in Θ'_t can be represented by a graph with $t - 1$ annotated nodes (with annotations $c_i \in \Omega_i$, $i = 1, \ldots, t-1$), plus at most t others annotated nodes with annotations in Ω_t, and

at most $2t$ leaves. Therefore, we can show that the number of partitions in Θ'_t is less than $(4t)^{2t-1}(s_t)^t\Pi_{i=1}^{t-1}s_i$ by counting the number of graphs with these properties. This is easily done since there are $(s_t)^t\Pi_{i=1}^{t-1}s_i$ different ways to annotate the annotated nodes ($\#\Omega_t = s_t$), and $(4t)^{2t-1}$ ways (for each of the two truth values) to dispose the edges outgoing from $2t-1$ annotated nodes and ingoing to $4t$ nodes.

Finally, since s_t increase for increasing t, we can write,

$$\#\bigcup_{t=1}^{T}\Theta'_t \leq (T+1)\Theta'_T \leq (4Ts_T)^{2T},$$

which completes the proof. □

Proof of Proposition 3. Throughout the proof we will assume that

$$\sup_{f\in\mathcal{F}[\tilde{\Theta}_T]} |\hat{\mathcal{E}}[f] - \mathcal{E}[f]| \leq q(m,\delta),$$

by Proposition 2, relaxing this assumption will just reduce by δ the confidence level of the final result.

For simplicity, let us fix an arbitrary t between 1 and T, and let us use the simplified notations $B := B_t$, $w := w_t$, $\mathcal{P} := \mathcal{P}_{t-1}$ and $\hat{\xi} := \hat{\xi}_{t-1}$.

Recall that the transformation $\hat{A}^B_{h,c}$, is a merge-split step. The parameter B defines the partition $\mathcal{B} = \{b_h\}_{h=1,\dots,2B} \prec \mathcal{P}$, with elements defined by

$$b_h = \{x \in X |\ M(h-1) < B\hat{f}_{\mathcal{P}}(x) + MB \leq Mh\},\ h = 2,\dots,2B, \tag{33}$$

the merge $\mathcal{P} \to \mathcal{P}^M$ is achieved replacing $\{\ell \in \mathcal{P} |\ \ell \subset b_h\}$ with the element $\{b_h\}$. The split $\mathcal{P}^M \to \mathcal{P}^{MS} := \hat{A}^B_{h,c}\mathcal{P}$ is achieved replacing $\{b_h\}$ with $\{b_h \cap c, b_h \cap \bar{c}\}$.

For the following analysis, it is useful observing that, for every $\mathbf{c} = (c_1, c_2, \dots, c_{2B}) \in (\Omega_t)^{2B}$, it holds

$$\sum_{h=1}^{2B} \hat{\mathcal{E}}[\mathcal{P}] - \hat{\mathcal{E}}[\hat{A}^B_{h,c_h}\mathcal{P}] = \hat{\mathcal{E}}[\mathcal{P}] - \hat{\mathcal{E}}[\mathcal{B}_{\mathbf{c}}] = \hat{\Delta}^M + \hat{\Delta}^S_{\mathbf{c}}, \tag{34}$$

where,

$$\hat{\Delta}^M := \hat{\mathcal{E}}[\mathcal{P}] - \hat{\mathcal{E}}[\mathcal{B}],\ \hat{\Delta}^S_{\mathbf{c}} := \hat{\mathcal{E}}[\mathcal{B}] - \hat{\mathcal{E}}[\mathcal{B}_{\mathbf{c}}],$$

and we introduced the partition $\mathcal{B}_{\mathbf{c}}$, obtained from $\mathcal{B}$ by the $2B$ splits, $\{b_h\} \to \{b_h \cap c_h, b_h \cap \bar{c}_h\}$, $h = 1,\dots,2B$. It is also important noticing that, since $\mathcal{P} \in \Theta_{t-1}$, it follows that $\mathcal{B}_{\mathbf{c}} \in \Theta'_t \subset \tilde{\Theta}_T$.

Now, we want to express the empirical split contribution $\hat{\Delta}^S_{\mathbf{c}}$ in terms of ideal estimators $f_{\mathcal{P}}$, defined by

$$f_{\mathcal{P}}(x) = \frac{1}{\rho_X(\ell(x))}\int_{\ell(x)} f_\rho\ d\rho_X,$$

with $x \in \ell(x) \in \mathcal{P}$, and ρ_X the marginal distribution of ρ over X.

In fact, observing that $\hat{\mathcal{E}}[\mathcal{B}_{\mathbf{c}}] := \hat{\mathcal{E}}[\hat{f}_{\mathcal{B}_{\mathbf{c}}}] \leq \hat{\mathcal{E}}[f_{\mathcal{B}_{\mathbf{c}}}]$ and $\mathcal{E}[\mathcal{B}] := \mathcal{E}[f_{\mathcal{B}}] \leq \mathcal{E}[\hat{f}_{\mathcal{B}}]$, we get

$$\begin{aligned} \hat{\Delta}^S_{\mathbf{c}} &= \hat{\mathcal{E}}[\mathcal{B}] - \hat{\mathcal{E}}[\mathcal{B}_{\mathbf{c}}] \geq \hat{\mathcal{E}}[\mathcal{B}] - \hat{\mathcal{E}}[f_{\mathcal{B}_{\mathbf{c}}}] \\ &\geq \mathcal{E}[\hat{f}_{\mathcal{B}}] - \mathcal{E}[\mathcal{B}_{\mathbf{c}}] - 2q(m,\delta) \\ &\geq \mathcal{E}[\mathcal{B}] - \mathcal{E}[\mathcal{B}_{\mathbf{c}}] - 2q(m,\delta) =: \Delta^S_{\mathbf{c}} - 2q(m,\delta). \end{aligned} \tag{35}$$

We also need the following estimates which follow from Propositions A3 and A2 . Since, by Eq. (33), the excursion of $\hat{f}_{\mathcal{P}}$ over b_h is not larger than M/B,

$$\hat{\Delta}^M = \hat{\mathcal{E}}[\mathcal{P}] - \hat{\mathcal{E}}[\mathcal{B}] \geq -\sum_{h=1}^{2B} \hat{\rho}_X(b_h) M^2 B^{-2} = -M^2 B^{-2}, \tag{36}$$

and,

$$\begin{aligned} \Delta^S_{\mathbf{c}} &= \mathcal{E}[\mathcal{B}] - \mathcal{E}[\mathcal{B}_{\mathbf{c}}] = \sum_{h=1}^{2B} \rho_X(b_h) \frac{\mathrm{cov}^2_{b_h}(f_\rho, \chi_{c_h})}{\mathrm{var}_{b_h}(\chi_{c_h})} \\ &\geq 4 \sum_{h=1}^{2B} \rho_X(b_h) \mathrm{cov}^2_{b_h}(f_\rho, \chi_{c_h}), \end{aligned} \tag{37}$$

where the $\mathrm{var}_\ell(\cdot)$ and $\mathrm{cov}_\ell(\cdot)$ are variance and covariance with respect to the probability measure over ℓ, $\rho_\ell(\cdot) = \rho_X(\cdot)/\rho_X(\ell)$.

Using Equations (34), (35), (36), we obtain,

$$\begin{aligned} & \hat{\mathcal{E}}[\mathcal{P}] - \min_{(h,c)\in\{1,\dots,2B\}\times\Omega_t} \hat{\mathcal{E}}[\hat{A}^B_{h,c}\mathcal{P}] \\ \geq\ & \hat{\mathcal{E}}[\mathcal{P}] - \frac{1}{2B}\sum_{h=1}^{2B} \min_{c\in\Omega_t} \hat{\mathcal{E}}[\hat{A}^B_{h,c}\mathcal{P}] \\ =\ & \frac{1}{2B} \max_{\mathbf{c}\in\Omega_t^{2B}} \left[\hat{\mathcal{E}}[\mathcal{P}] - \hat{\mathcal{E}}[\mathcal{B}_{\mathbf{c}}]\right] = \frac{1}{2B}\left(\hat{\Delta}^M + \max_{\mathbf{c}\in\Omega_t^{2B}} \hat{\Delta}^S_{\mathbf{c}}\right) \\ \geq\ & \frac{1}{2B}\left(\max_{\mathbf{c}\in\Omega_t^{2B}} \Delta^S_{\mathbf{c}} - M^2B^{-2} - 2q(m,\delta)\right). \end{aligned} \tag{38}$$

We now want to prove that, letting $\delta_t := \frac{6\delta}{\pi^2 t^2}$, with probability greater than $1-\delta_t$, the inequality

$$\begin{aligned} & \max_{\mathbf{c}\in\Omega_t^{2B}} \Delta^S_{\mathbf{c}} \\ \geq\ & V^{-2}\hat{\xi}^2 - (w + 2q(m,\delta)V^{-2})\hat{\xi} - (M^2B^{-2} + q(m,\delta))w - 3V^{-2}q^2(m,\delta), \end{aligned} \tag{39}$$

holds. This result together with equation (38), observing that $\sum_t \delta_t \leq \delta$, will complete the proof.

In order to prove eq. (39), first notice that by eq. (37),

$$\max_{\mathbf{c}\in\Omega_t^{2B}} \Delta^S_{\mathbf{c}} = \sum_{h=1}^{2B} \max_{c\in\Omega_t} \Delta^S_h(c), \tag{40}$$

with,

$$\Delta^S_h(c) := \frac{\rho_X(b_h)\mathrm{cov}^2_{b_h}(f_\rho, \chi_c)}{\mathrm{var}_{b_h}(\chi_c)}.$$

Since,

$$0 \leq \Delta^S_h(c) \leq \rho_X(b_h)\mathrm{var}_{b_h}(f_\rho), \text{ for all } c \in \mathcal{C}$$

applying Markov inequality to the positive random variable $\rho_X(b_h)\mathrm{var}_{b_h}(f_\rho) - \Delta_h^S$, after a simple algebraic computation we obtain

$$\begin{aligned} &\Pr_{c\sim\pi}\left(\Delta_h^S(c) \le \rho_X(b_h)\mathrm{var}_{b_h}(f_\rho) - (\rho_X(b_h)\mathrm{var}_{b_h}(f_\rho) - \mathbb{E}\Delta_h^S(c))(1+w)\right) \\ \le\ &\left(\frac{\delta_t}{2B}\right)^{\frac{1}{s_t}}, \end{aligned} \tag{41}$$

which implies,

$$\begin{aligned} &\Pr_{\Omega_t\sim\pi^{s_t}}\left(\forall h\in\{1,\dots,2B\}\ \max_{c\in\Omega_t}\Delta_h^S(c) \ge -w\rho_X(b_h)\mathrm{var}_{b_h}(f_\rho) + \mathbb{E}\Delta_h^S(c)(1+w)\right) \\ \ge\ &1-\delta_t. \end{aligned} \tag{42}$$

Hence, from eq. (40), eq. (37), eq. (42), Proposition 25, Hypothesis 3 and the convexity of x^2, with probability greater than $1-\delta_t$, it holds

$$\begin{aligned} \max_{\mathbf{c}\in\Omega_t^{2B}}\Delta_{\mathbf{c}}^S &\ge \sum_{h=1}^{2B}\left(-w\rho_X(b_h)\mathrm{var}_{b_h}(f_\rho) + \mathbb{E}\Delta_h^S(c)(1+w)\right) \\ &= w(\mathcal{E}[f_\rho] - \mathcal{E}[\mathcal{B}]) + (1+w)\mathbb{E}\Delta_{\mathbf{c}}^S \\ &\ge w(\mathcal{E}[f_\rho] - \hat{\mathcal{E}}[\mathcal{B}] - q(m,\delta)) + V^{-2}\sum_{h=1}^{2B}\rho_X(b_h)\mathrm{var}_{b_h}^2(f_\rho) \\ &\ge w(\mathcal{E}[f_\rho] - \hat{\mathcal{E}}[\mathcal{P}] - M^2B^{-2} - q(m,\delta)) + V^{-2}(\mathcal{E}[\mathcal{P}] - \mathcal{E}[f_\rho])^2 \\ &\ge w(-\hat{\xi} - M^2B^{-2} - q(m,\delta)) + V^{-2}(\hat{\xi}^2 - 2q(m,\delta)\hat{\xi} - 3q^2(m,\delta)), \end{aligned} \tag{43}$$

where the last inequality is obtained observing that

$$\begin{aligned} (\mathcal{E}[\mathcal{P}] - \mathcal{E}[f_\rho])^2 &\ge (\hat{\xi} + q(m,\delta) - 2q(m,\delta))(\mathcal{E}[\mathcal{P}] - \mathcal{E}[f_\rho]) \\ &\ge \hat{\xi}^2 - q^2(m,\delta) - 2q(m,\delta)(\mathcal{E}[\mathcal{P}] - \mathcal{E}[f_\rho]) \\ &\ge \hat{\xi}^2 - 2q(m,\delta)\hat{\xi} - 3q^2(m,\delta). \end{aligned} \tag{44}$$

As claimed, eq. (43) implies eq. (39), which completes the proof. □

Proposition A2. *Let $\mathcal{P}'$ be a partition of X, and $\mathcal{P}$ the partition obtained from $\mathcal{P}'$ by the split $\{b\}\to\{\ell,\bar{\ell}\}$, with $\ell = b\cap c$ and $\bar{\ell} = b\cap\bar{c}$, for some $b\in\mathcal{P}'$ and subset c. Then, for every ρ,it holds*

$$\mathcal{E}[\mathcal{P}'] - \mathcal{E}[\mathcal{P}] = \rho_X(b)\frac{\mathrm{cov}_b^2(f_\rho,\chi_c)}{\mathrm{var}_b(\chi_c)},$$

where $\mathrm{var}_b(\cdot)$ and $\mathrm{cov}_b(\cdot)$ are relative to the probability measure on b obtained by restricting ρ_X.

Proof. The Proposition follows by direct computation. In fact it is easy to verify that

$$\begin{aligned}\mathcal{E}[\mathcal{P}'] - \mathcal{E}[\mathcal{P}] &= \frac{\rho_X(\ell)\rho_X(\bar{\ell})(f_{\mathcal{P}}(\ell) - f_{\mathcal{P}}(\bar{\ell}))^2}{\rho_X(b)},\\ \mathrm{cov}_b(f_\rho, \chi_c) &= \frac{\rho_X(\ell)\rho_X(\bar{\ell})(f_{\mathcal{P}}(\ell) - f_{\mathcal{P}}(\bar{\ell}))}{\rho_X^2(b)},\\ \mathrm{var}_b(\chi_c) &= \frac{\rho_X(\ell)\rho_X(\bar{\ell})}{\rho_X^2(b)},\end{aligned}$$

where $f_{\mathcal{P}}(\ell)$ is the value of $f_{\mathcal{P}}(x)$ for $x \in \ell$.

□

Proposition A3. *Let $\mathcal{P}$ be a partition of X, and $\mathcal{P}'$ the partition obtained from $\mathcal{P}$ by the merge $\{\ell_1, \ldots, \ell_k\} \to \{b\}$, with $b = \ell_1 \cup \cdots \cup \ell_k$, for some $\ell_1, \ldots, \ell_k \in \mathcal{P}$. Then, for every ρ,it holds*

$$\mathcal{E}[\mathcal{P}'] - \mathcal{E}[\mathcal{P}] \leq \rho_X(b)\left(\max_{x\in b} f_{\mathcal{P}}(x) - \min_{x\in b} f_{\mathcal{P}}(x)\right)^2.$$

Proof. From the proof of Proposition A2 we get

$$\mathcal{E}[\mathcal{P}'] - \mathcal{E}[\mathcal{P}] = \frac{\rho_X(\ell)\rho_X(\bar{\ell})(f_{\mathcal{P}}(\ell) - f_{\mathcal{P}}(\bar{\ell}))^2}{\rho_X(b)} \leq \rho_X(\bar{\ell})(f_{\mathcal{P}}(\ell) - f_{\mathcal{P}}(\bar{\ell}))^2.$$

The Proposition can be proved by induction on k, letting $\ell = \ell_1 \cup \cdots \cup \ell_{k-1}$ and $\bar{\ell} = \ell_k$.

□

Proof of Proposition 4. It is convenient introducing the sequence of functions $\phi_t : \mathbb{R}^+ \to \mathbb{R}$, for $t \geq 0$, defined by

$$\phi_t(x) = \begin{cases} x - x^2 & \text{if } x \geq \psi(2t),\\ x + \psi(2t)^2 & \text{if } x < \psi(2t).\end{cases}$$

We want to prove that, if $e_2 \leq e/(1+3e)$, and for every $t \geq 3$, it is true that $e_{t+1} \leq e_t - 2e_t^2 + 1/2 e_t\psi_1(2t) + 1/2\psi_2(2t)^2$ and $0 \leq \psi_i(t) \leq \psi(t)$, then $e_t \leq \psi(t) := a_t$. Hence, the Proposition will follow by the renaming $e_{t+3} \to e_t$.

We proceed by proving that for every $t \geq 3$, the following three statements hold,

$$e_{t+1} \leq \phi_t(e_t), \tag{45}$$

$$e_t \leq a_t \Rightarrow \phi_t(e_t) \leq \phi_t(a_t), \tag{46}$$

$$\phi_t(a_t) \leq a_{t+1}. \tag{47}$$

The claimed result will follow from the three relations above, by induction on t. In fact, by assumption $e_3 \leq e/(1+3e) = a_3$, and if $e_t \leq a_t$ for $t \geq 3$, chaining the relations above, we get $e_{t+1} \leq \phi_t(e_t) \leq \phi_t(e_t) \leq \phi_t(a_t) \leq \phi_t(a_t) \leq a_{t+1}$.

Inequality (45) can be proved observing that, since $0 \leq \psi_i \leq \psi$,

$$e_{t+1} \leq e_t - 2e_t^2 + \frac{1}{2}e_t\psi_1(2t) + \frac{1}{2}\psi_2(2t)^2 \leq \begin{cases} e_t - e_t^2 & \text{if } e_t \geq \psi(2t),\\ e_t + \psi(2t)^2 & \text{if } e_t < \psi(2t).\end{cases}$$

In order to prove relation (46), we treat separately the two cases, $e_t \geq \psi(2t)$ and $e_t < \psi(2t)$.

Case $e_t \geq \psi(2t)$. Since $a_t = \psi(t) \geq \psi(2t)$, by the definition of ϕ_t, we have to prove

$$e_t - e_t^2 \leq a_t - a_t^2.$$

The inequality above is true, since the function $x - x^2$ is monotonic increasing in for $x \leq 1/2$, and $e_t \leq a_t \leq e/(1+3e) < 1/2$.

Case $e_t < \psi(2t)$. Recalling again that $a_t = \psi(t)$, by the definition of ϕ_t, we get

$$\begin{aligned} \phi_t(a_t) - \phi_t(e_t) &= a_t - a_t^2 - e_t - \psi(2t)^2 \geq \psi(t) - \psi(2t) - \psi(t)^2 - \psi(2t)^2 \\ &= e^4 t^2 (2t-5) + e^2(3et+1)(t-2). \end{aligned}$$

This proves the statement $\phi_t(a_t) - \phi_t(e_t) \geq 0$, since by assumption $t \geq 3$.

Finally, inequality (47) can be proved observing that, since $a_t \geq \psi(2t)$, it holds $\phi_t(a_t) = a_t - a_t^2$, and one can write,

$$\begin{aligned} \phi(a_t) &= \frac{e(1+et-e)}{(1+et)^2} \\ &= a_{t+1}\frac{(1+et-e)(1+et+e)}{(1+et)^2} = a_{t+1}\frac{(1+et)^2 - e^2}{(1+et)^2} \leq a_{t+1}. \end{aligned}$$

□

Proof of Proposition 5. First, observe that the constraints (19) are straightforwardly implied by the set of inequalities

$$w_t \leq \frac{1}{2} B_t \psi_t, \tag{48}$$

$$q(m,\delta) \leq \frac{1}{4} V^2 B_t \psi_t, \tag{49}$$

$$(1+w_t) M^2 B_t^{-4} \leq 2\max(M^2, V^2)\psi_t^2, \tag{50}$$

$$(2 + w_t + 3V^{-2} q(m,\delta)) q(m,\delta) \leq 2\max(M^2, V^2) B_t^2 \psi_t^2. \tag{51}$$

We now proceed to the proof of the inequalities above.

Step 1. Proof of (48).First, observe that

$$\begin{aligned} s_t &\geq 4\psi_t^{-\frac{1}{2}} \log \frac{6\psi_t^{-\frac{5}{2}}}{5\delta} \\ \left(\frac{6}{5} \geq \frac{\pi^2(1+1/\sqrt{5})}{12}\right) &\geq 4\psi_t^{-\frac{1}{2}} \log \frac{\pi^2(1+1/\sqrt{5})\psi_t^{-\frac{5}{2}}}{12\delta} \\ \left(\psi_t^{-\frac{1}{2}} \geq \sqrt{5}\right) &\geq 4\psi_t^{-\frac{1}{2}} \log \frac{\pi^2(1+\psi_t^{-\frac{1}{2}})\psi_t^{-2}}{12\delta} \\ \left(\psi_t^{-\frac{1}{2}} \leq B_t \leq \psi_t^{-\frac{1}{2}} + 1\right) &\geq 4B_t^{-1}\psi_t^{-1} \log \frac{\pi^2 t^2 B_t}{3\delta}. \end{aligned} \tag{52}$$

Second, $w_t \leq 1/2$, since by eq. (52),

$$\log(1+w_t) = \frac{1}{s_t} \log \frac{\pi^2 t^2 B_t}{3\delta} \leq \frac{1}{4} B_t \psi_t \leq \frac{1}{4}. \tag{53}$$

Hence,

$$\begin{aligned} (x \leq (1+\bar{x})\log(1+x) \text{ for } x \leq \bar{x}) \; w_t &\leq \frac{3}{2}\log(1+w_t) = \frac{3}{2}\frac{1}{s_t}\log\frac{\pi^2 t^2 B_t}{3\delta} \\ (\text{eq. (52)}) &\leq \frac{1}{2} B_t \psi_t. \end{aligned}$$

Step 2. Proof of (49). It follows directly from eq. (18) and $B_t \geq 1$.

Step 3. Proof of (50). Using eq. (53) and $B_t \geq \psi_t^{-\frac{1}{2}}$, we get

$$(1 + w_t)M^2 B_t^{-4} \leq 2M^2 B_t^{-4} \leq 2\max(M^2, V^2)\psi_t^2.$$

Step 4. Proof of (51). From eq. (52) and eq. (18) we get,

$$w_t + 3V^{-2}q(m, \delta) \leq \frac{1}{2} + \frac{3}{20} \leq 1.$$

Therefore, using again $\psi_t^{-\frac{1}{2}} \leq B_t$,

$$(2w_t + 3V^{-2}q(m, \delta))q(m, \delta) \leq 3q(m, \delta) \leq \frac{3}{4}V^2\psi_t \leq 2\max(M^2, V^2)B_t^2\psi_t^2,$$

which concludes the proof of eq. (51).

We complete the proof of the Proposition showing that, for $m \geq \left(\frac{V}{4M}\right)^2 + e$, eq. (8) implies eq. (18). First, observe that introducing the constant $\beta := (\frac{V}{4M})^4 \frac{m}{\delta \log m}$, we get

$$\begin{aligned} \text{(eq. (8))} \ \frac{4}{\delta\psi_T^3} &\leq \beta\frac{\log m}{4\log^2(m/\delta)} \leq \beta\log^{-1}(m^4/\delta^4) \\ \left(m \geq \left(\frac{V}{4M}\right)^2 + e\right) &\leq \beta\log^{-1}(m^3/\delta + e) \leq \beta\log^{-1}(\beta + e). \end{aligned} \tag{54}$$

Second, using the inequality $\psi_T^{-1} \geq 5$, we obtain

$$\begin{aligned} s_T &= \left\lceil 4\psi_T^{-\frac{1}{2}} \log\frac{6\psi_T^{-\frac{5}{2}}}{5\delta} \right\rceil \\ \left(\log x \leq \frac{\log\bar{x}}{\bar{x}}x \text{ for } x \geq \bar{x} \geq 0.\ \frac{6\psi_T^{-\frac{5}{2}}}{5\delta} \geq 64\right) &\leq \left\lceil \frac{1}{2\delta\psi_T^3} \right\rceil \leq \frac{1}{2\delta\psi_T^3} + 1 \\ \left(x + 1 \leq (1 + \frac{1}{\bar{x}})x \text{ for } x \geq \bar{x} \geq 0.\ \frac{1}{2\delta\psi_T^3} \geq 5\right) &\leq \frac{3}{5\delta\psi_T^3}. \end{aligned} \tag{55}$$

Finally, introducing the constant $\alpha := 2T\log(8mTs_T) + \log(2/\delta)$, we conclude the proof,

$$
\begin{aligned}
& & \left(\frac{4q(m,\delta)}{V^2\psi_T}\right)^2 \\
\text{(eq. (10))} & \leq & \left(\frac{16M^2}{V^2\psi_T\sqrt{m}}(\sqrt{\alpha}+\sqrt{m^{-1}})\right)^2 \\
\left(\sqrt{x}+\sqrt{y}\leq\sqrt{2(x+y)}\right) & \leq & \left(\frac{4M}{V}\right)^4\frac{2}{\psi_T^2 m}(\alpha+m^{-1}) \\
(m\geq e) & \leq & \left(\frac{4M}{V}\right)^4\frac{2\log m}{\psi_T^2 m}(2T\log(8Ts_T)+\log(2e/\delta)) \\
& \leq & \left(\frac{4M}{V}\right)^4\frac{2\log m}{\psi_T^3 m}\log\frac{8es_T}{\delta\psi_T} \\
\text{(eq. (55))} & \leq & \left(\frac{4M}{V}\right)^4\frac{2\log m}{\psi_T^3 m}\log\frac{16}{\delta^2\psi_T^4} \\
& \leq & \frac{1}{\beta}\frac{4}{\delta\psi_T^3}\log\frac{4}{\delta\psi_T^3} \\
\text{(eq. (54))} & \leq & \frac{\log\beta-\log\log(\beta+e)}{\log(\beta+e)}\leq 1.
\end{aligned}
$$

□

Appendix B. Estimates over Euclidean spaces

In this Appendix we determine covariance estimates similar to those given in Theorem 2 while relaxing the assumption that the measure π over $\mathcal{C}$ is finite. In fact, while in the previous sections we assumed that π is a probability measure (i.e. $\pi(\mathcal{C})=1$), here we allow $\pi(\mathcal{C})=+\infty$. The main estimate is given in Theorem B2. The obvious application of this result is an extension of Example 1 from functions over balls in E^d to functions over E^d. We will prove Theorem B2, then we will apply it to that Example.

We first observe that the distance D, defined in (1), is *negative* ([9] Sec.6.2). Recall that a kernel $D: X\times X\to\mathbb{R}$ is said to be negative when it is symmetric and for arbitrary $x_1,\dots,x_n$ in X, and $r_1,\dots,r_n$ in $\mathbb{R}$ with

$$\sum_{i=1}^n r_i = 0,$$

fulfills

$$\sum_{i,j=1}^n r_i D(x_i,x_j) r_j \leq 0.$$

Proposition B1. *D(x,y) is a negative kernel on X.*

Proof. Symmetry was proved in Proposition 1. Assume $x_1, \ldots, x_n$ and $r_1, \ldots, r_n$ as in the definition above. By Definition 1

$$\begin{aligned}\sum_{i,j=1}^{n} r_i D(x_i, x_j) r_j &= \int_{\mathcal{C}} \sum_{i,j=1}^{n} r_i \left(\chi_c(x_i)(1-\chi_c(x_j)) + (1-\chi_c(x_i))\chi_c(x_j)\right) r_j d\pi(c) \\ &= -2 \int_{\mathcal{C}} \left(\sum_{i=1}^{n} r_i \phi_{(x_i)}(c) \right)^2 d\pi(c) \leq 0,\end{aligned}$$

which proves the proposition. □

The importance of negativity follows from the theorem by Schönberg, below ([9] Th.4).

Theorem B1. *$D(x, y)$ is a negative kernel iff $K_a(x, y) = e^{-aD(x,y)}$ is symmetric and positive-definite for every $a > 0$.*

From Schönberg theorem and Proposition B1 it follows that the kernels K_a are positive-definite.

Proposition B2. *For all $a > 0$, the kernel*

$$K_a(x, y) = e^{-aD(x,y)}$$

is a Mercer kernel over the metric space (X, D).

We will denote $\mathcal{H}_a$ the RKHS induced by K_a over X.

It is possible to establish a variance estimate for functions f belonging to the RKHSs $\mathcal{H}_a$, analogous to the result stated in Theorem 2.

Theorem B2. *Let $\epsilon > 0$ and $f \in \mathcal{H}_a$ for all $a < \epsilon$, moreover assume*

$$V^2(f) := \frac{1}{2} \limsup_{a \to 0} \; a \|f\|^2_{\mathcal{H}_a} < +\infty,$$

then the following inequality holds

$$\mathrm{var}^2_\nu(f) \leq 4V^2(f) \int_{\mathcal{C}} \mathrm{cov}^2_\nu(f, \chi_c) d\pi(c). \tag{56}$$

Proof. We first prove that for every $a > 0$ and $f \in \mathcal{H}_a$ it holds

$$\mathrm{var}^2(f) \leq a \|f\|^2_{\mathcal{H}_a} \int_{\mathcal{C}} \mathrm{cov}^2(f, \chi_c) d\pi(c) + O(a^2 \|f\|^2_{\mathcal{H}_a}), \tag{57}$$

hence the Theorem will follow letting a going to zero in (57).

Reasoning as in the proof of Theorem 2 we get

$$\mathrm{var}^2(f) \leq \|f\|^2_{\mathcal{H}_a} \langle L_{K_a} P_0 f, P_0 f \rangle_\nu . \tag{58}$$

Now, using definition (22) and reasoning again as in Theorem 2, we obtain

$$\begin{aligned}&\langle L_{K_a} P_0 f, P_0 f \rangle_\nu \\ &= \langle \mathbf{1}, P_0 f \rangle^2_\nu - a \int_{X \times X} (P_0 f)(x) D(x, y) (P_0 f)(y) d\nu^2(x, y) + O(a^2) \\ &= 2a \int_{\mathcal{C}} \mathrm{cov}^2(f, \chi_c) \; d\pi(c) + O(a^2).\end{aligned} \tag{59}$$

Equation (57) follows from (58) and (59). □

The estimate (62) obtained in Example 1′ below, is used in Section 3 to give an alternate analysis of Example 1.

Example 1′ Let $X = E^d$. Define

$$\mathcal{C} := \{c(\omega, p)|\ (\omega, p) \in S^{d-1} \times \mathbb{R}\},$$

where

$$c(\omega, p) := \{x \in X|\ \omega \cdot (o - x) > p\},$$

for some fixed point o.

Endow $\mathcal{C}$ with the σ-field $\mathcal{F}$ and the measure π induced by the natural product measure over $S^{d-1} \times \mathbb{R}$. Clearly Hypothesis 1 is fulfilled but, $\pi(\mathcal{C}) = +\infty$.

As in Example 1, the kernel D is given by

$$D(x, y) = C(d) \|x - y\|,$$

where

$$C(d) := \frac{4|S^{d-2}|}{d-1}.$$

We now want to compute explicitly the RKHS norm $\|f\|_{\mathcal{H}_a}$ and the complexity function $V^2(f)$. Since equation (24) holds for every non-degenerate measure ν over X, we can identify it with the Lebesgue measure μ. The operator $L_{K_a} : \mathcal{L}^2(E^d, \mu) \to \mathcal{L}^2(E^d, \mu)$ is the convolution operator with kernel

$$K_a(x) = e^{-aC(d)\|x\|}.$$

Recalling the relation between convolution product and Fourier transform

$$\widehat{L_{K_a} f} = \widehat{K_a}\ \widehat{f},$$

where

$$\widehat{f}(\xi) = \int f(x) e^{-\xi \cdot x} d\mu(x),$$

and Parseval's theorem, that is

$$\langle f, g \rangle_\mu = \frac{1}{(2\pi)^d} \left\langle \widehat{f}, \widehat{g} \right\rangle_\mu,\ \forall f, g \in \mathcal{L}^2(E^d, \mu),$$

we obtain

$$\left\| L_{K_a}^{-\frac{1}{2}} f \right\|_\mu^2 = \left\langle f, L_{K_a}^{-1} f \right\rangle_\mu = \frac{1}{(2\pi)^d} \left\langle \hat{f}, (\widehat{K_a})^{-1} \hat{f} \right\rangle_\mu.$$

Since the expression for the Fourier transform of K_a is [8]

$$\widehat{K_a}(\xi) = \frac{2aC(d)(2\pi)^d}{|S^d|} (a^2 C(d)^2 + \|\xi\|^2)^{-\frac{d+1}{2}},$$

we obtain

$$\begin{aligned} a \|f\|_{\mathcal{H}_a}^2 &= \frac{\pi}{2(2\pi)^d} \int (a^2 C(d)^2 + \|\xi\|^2)^{\frac{d+1}{2}} |\widehat{f}(\xi)|^2 d\mu(\xi) \\ &= \frac{\pi}{2} \left\langle f, \left(a^2 C(d)^2 + (-\triangle)^{\frac{d+1}{2}} \right) f \right\rangle_\mu. \end{aligned}$$

From the relation above it is clear that for all $a > 0$, the space $\mathcal{H}_a$ is equal to

$$H^{\frac{d+1}{2}}(E^d) := \{f \in \mathcal{L}^2(E^d, \mu)|\ \left\langle f, (-\triangle)^{\frac{d+1}{2}} f \right\rangle_\mu < +\infty\}, \tag{60}$$

and for every $f \in \mathcal{H}$ the function $\phi(a) = a \|f\|_{\mathcal{H}_a}^2$ is non-decreasing.

From the above observations it follows that the complexity function $V^2(f)$ appearing in the text of Theorem B2 is given by

$$V^2(f) = \frac{\pi}{4} \left\langle f, (-\triangle)^{\frac{d+1}{2}} f \right\rangle_{\mathcal{L}^2(E^d,\mu)}, \tag{61}$$

for every $f \in H^{\frac{d+1}{2}}(E^d)$.

Therefore by Theorem B2 we obtain

$$\int_{\mathcal{C}} \mathrm{cov}^2(f, \chi_c) d\pi(c) \geq \frac{\mathrm{var}^2(f)}{\pi \left\langle f, (-\triangle)^{\frac{d+1}{2}} f \right\rangle_{\mathcal{L}^2(E^d,\mu)}}. \tag{62}$$

Appendix C. Estimates over spheres

Let $\mathcal{L}^2(S^{d-1})$ be the space of square-integrable functions on the $(d-1)$-dimensional sphere endowed with the natural measure dS^{d-1}. We consider the kernel K on $\mathcal{L}^2(S^{d-1})$ obtained by equalities (22) and (23) starting from the set $\mathcal{C}$ of hemispheres of S^{d-1} endowed with uniform measure (see Example 1 in the text). The following result characterizes the RKHS induced by K. It is a direct corollary of Proposition C4. Here, $\triangle$ is the Laplace-Beltrami operator on S^{d-1} (see for example [15] §14.18).

Theorem C1. *The RKHS $\mathcal{H}$ induced by the kernel $K(x,y) = 1 - \frac{2}{\pi} d(x,y)$ on the sphere S^{d-1} for even $d \geq 4$ is characterized as follows*

$$\mathcal{H} = \{ f \in \mathbb{P}_A \mathcal{L}^2(S^{d-1}) | \ \left\langle f, P_{d/2}(-\triangle) f \right\rangle_{\mathcal{L}^2(S^{d-1})} < +\infty \},$$

where $\mathbb{P}_A$ is the orthogonal projector over the subspace of antisymmetric functions in $\mathcal{L}^2(S^{d-1})$ and $P_{d/2}$ is the polynomial

$$P_{d/2}(z) := \prod_{i=1}^{d/2} (z + (d-2i)(2i-2)). \tag{63}$$

Moreover for every $f \in \mathcal{H}$,

$$\|f\|^2_{\mathcal{H}} = C \left\langle f, P_{d/2}(-\triangle) f \right\rangle_{\mathcal{L}^2(S^{d-1})},$$

where

$$C := \frac{\pi^2}{2} \left((4\pi)^{d/2} \Gamma(d/2) \right)^{-1}. \tag{64}$$

We now proceed to the proof of Proposition C4. We first need some preliminary technical results.

Proposition C1. *For any positive odd integer l and $d > 2$, it holds*

$$\frac{d^{l-1}}{dt^{l-1}} (1-t^2)^{l+\frac{d-3}{2}} \Big|_{t=0} = (-1)^{\frac{l-1}{2}} \frac{\Gamma(l)\Gamma(l+\frac{d-1}{2})}{\Gamma(\frac{l+1}{2})\Gamma(\frac{l+d}{2})}.$$

Proof. The proposition follows from the identity

$$\left(\frac{d}{dt}\right)^l (1-t^2)^{l+\frac{d-3}{2}} = l!\Gamma(l+\frac{d-1}{2}) \sum_{i=0}^{[\frac{l}{2}]} \frac{(-1)^i (1-t^2)^{i+\frac{d-3}{2}} (-2t)^{l-2i}}{i!(l-2i)!\Gamma(i+\frac{d-1}{2})},$$

where $[x]$ stands for the integer part of x. The expression above can be verified substituting the explicit form of Legendre harmonics ([15] §2.32) into the Rodrigues

representation ([15] §2 Lemma 4). The proof is completed substituting d, l and t with $d+2$, $l-1$ and 0 respectively. □

Proposition C2. *For any positive integer l and even positive integer d, it holds*

$$\left(\frac{(l+d-2)!!}{(l-2)!!}\right)^2 = P_{d/2}(l(l+d-2)),$$

where $P_{d/2}$ is the polynomial defined in (63).

Proof. Since

$$\frac{(l+d-2)!!}{(l-2)!!} = \prod_{i=1}^{d/2}(l+2i-2) = \prod_{i=1}^{d/2}(l-2i+d),$$

it holds

$$\left(\frac{(l+d-2)!!}{(l-2)!!}\right)^2 = \prod_{i=1}^{d/2}(l+2i-2)(l-2i+d).$$

The proposition follows from the identity above observing that

$$(l+2i-2)(l-2i+d) = l(l+d-2) + (d-2i)(2i-2).$$

□

In the following we denote by $\mathcal{Y}_l(d)$, $l = 0, 1, \ldots$, the space of spherical harmonics of degree l on S^{d-1}. Moreover we fix its orthonormal basis

$$\{Y_l^m | m = 0, \ldots, N(d,l)\},$$

with the constraint that Y_l^0 be the zonal spherical harmonic relative to the subgroup $SO(d-1)$ of rotations around the axis defined by an arbitrarily fixed point o on S^{d-1} ([16], Chapter IX, §3.1). It is known that there is exactly one such normalized function in $\mathcal{Y}_l(d)$. It can be expressed in terms of the Gegenbauer polynomials $C_l^m(t)$ ([16], Chapter IX, §3.5, eq.(7))

$$Y_l^0(x) = \frac{2l+d-2}{(d-2)\sqrt{|S^{d-1}|N(d,l)}} C_l^{(d-2)/2}(x \cdot o), \tag{65}$$

where the eigenspace dimension $N(d,l)$ is given by([16], Chapter IX, §2.5, eq.(11))

$$N(d,l) = \frac{\Gamma(d+l-2)(2l+d-2)}{\Gamma(d-1)l!}.$$

Proposition C3. *If $f \in \mathcal{Y}_l(d)$, with l=0,1,... and $d > 2$ then*

$$L_K f = \lambda_l f,$$

where

$$\lambda_l = (1-(-1)^l)\ 2^{-d-1}\Gamma(d/2)^{-2}C^{-1}\ B(l/2, d/2)^2, \tag{66}$$

with C the constant introduced in the text of Theorem C1.

Proof. By equalities (22) and (23), the kernel K is given by

$$K(x,y) = \frac{1}{|S^{d-1}|}\int_{S^{d-1}} \sigma(x \cdot p)\sigma(y \cdot p) dS^{d-1}(p),$$

where we identified the points of the sphere with vectors in the Euclidean d-dimensional space E^d, and σ is the sign function.

Now, let us consider the group of rotations of E^d, $G := SO(d)$. Let $d\mathbf{g}$ be its normalized invariant measure. By the factorization of $SO(d)$ in S^{d-1} times $SO(d-1)$ (see [16], Chapter IX, §1.4), the expression above can be rewritten as follows

$$K(x,y) = \int_G h(\mathbf{g}^{-1}x)h(\mathbf{g}^{-1}y)d\mathbf{g},$$

where

$$h(x) := \sigma(x \cdot o).$$

Since h is zonal spherical relative to the pole o, it admits the expansion

$$h = \sum_{l=0}^{+\infty} c_l Y_l^0.$$

We want to evaluate the eigensystem of the integral operator $L_K : \mathcal{L}^2(S^{d-1}) \to \mathcal{L}^2(S^{d-1})$ defined by the kernel K and the measure dS^{d-1}. It is clear that $\mathcal{Y}_l(d)$ are the eigenspaces of L_K, therefore we are left with computing the corresponding eigenvalues λ_l.

Recalling the transformation properties of the harmonics in $\mathcal{Y}_l(d)$ under the action of the group $SO(d)$ ([16], Chapter IX, §4.2, eq.(1) and §4.1, eq.(5)), we can write

$$\begin{aligned} \lambda_l &= \left\langle Y_l^0, L_K Y_l^0 \right\rangle_{\mathcal{L}^2(S^{d-1})} = |c_l|^2 \int_G \left| \int_{S^{d-1}} \overline{Y_l^0(x)} Y_l^0(\mathbf{g}^{-\mathbf{1}}x) dS^{d-1} \right|^2 d\mathbf{g} \\ &= |c_l|^2 \int_G |t_{0,0}^{d,l}(\mathbf{g})|^2 d\mathbf{g} = \frac{|S^{d-1}||c_l|^2}{N(d,l)}. \end{aligned} \tag{67}$$

The coefficients c_l can be computed using equation (65)

$$\begin{aligned} c_l &= \left\langle h, Y_l^0 \right\rangle_{\mathcal{L}^2(S^{d-1})} \\ &= \frac{2l+d-2}{(d-2)\sqrt{|S^{d-1}|N(d,l)}} \int_{S^{d-1}} \sigma(x \cdot o) C_l^{(d-2)/2}(x \cdot o) dS^{d-1} \\ &= \frac{(2l+d-2)|S^{d-2}|}{(d-2)\sqrt{|S^{d-1}|N(d,l)}} \int_0^{\pi} (\sin\theta)^{d-2} \sigma(\cos\theta) C_l^{(d-2)/2}(\cos\theta) d\theta \\ &= \frac{(2l+d-2)|S^{d-2}|}{(d-2)\sqrt{|S^{d-1}|N(d,l)}} \int_{-1}^{+1} \sigma(t) C_l^{(d-2)/2}(t)(1-t^2)^{(d-3)/2} dt. \end{aligned} \tag{68}$$

Using identity in [16], Chapter IX, §4.8, eq.(8), integrating by parts $l-1$ times, and using Proposition C1 we get

$$
\begin{aligned}
(69)\qquad & \int_{-1}^{+1} \sigma(t) C_l^{(d-2)/2}(t)(1-t^2)^{(d-3)/2} dt \\
= \; & \frac{\Gamma(d+l-2)\Gamma(\frac{d-1}{2})}{2(-2)^{l-1} l! \Gamma(d-2)\Gamma(l+\frac{d-1}{2})} \int_{-1}^{+1} \frac{d}{dt}\left(\sigma(t)\right) \frac{d^{l-1}}{dt^{l-1}} \left((1-t^2)^{l+(d-3)/2}\right) dt \\
= \; & \frac{\Gamma(d+l-2)\Gamma(\frac{d-1}{2})}{(-2)^{l-1} l! \Gamma(d-2)\Gamma(l+\frac{d-1}{2})} \frac{d^{l-1}}{dt^{l-1}} \left.(1-t^2)^{l+(d-3)/2}\right|_{t=0} \\
= \; & \frac{1-(-1)^l}{2} \frac{\Gamma(l)\Gamma(d+l-2)\Gamma(\frac{d-1}{2})}{(-4)^{\frac{l-1}{2}} l! \Gamma(d-2)\Gamma(\frac{l+1}{2})\Gamma(\frac{l+d}{2})}.
\end{aligned}
$$

By equations (67), (68) and (69), finally, for odd l, we get

$$
\begin{aligned}
\sqrt{\lambda_l} &= N(d,l)^{-1/2}|c_l| \\
&= \frac{(2l+d-2)|S^{d-2}|}{\sqrt{|S^{d-1}|}(d-2)N(d,l)} \left| \int_{-1}^{+1} \sigma(t) C_l^{(d-2)/2}(t)(1-t^2)^{(d-3)/2} dt \right| \\
&= \frac{|S^{d-2}|\Gamma(\frac{d-1}{2})\Gamma(l)}{\sqrt{|S^{d-1}|}2^{l-1}\Gamma(\frac{l+1}{2})\Gamma(\frac{l+d}{2})} = \frac{|S^{d-2}|\Gamma(\frac{d-1}{2})\Gamma(\frac{l}{2})}{\sqrt{\pi|S^{d-1}|}\Gamma(\frac{l+d}{2})} = \frac{\sqrt{|S^{d-1}|}}{\pi} B(\frac{l}{2},\frac{d}{2},),
\end{aligned}
$$

where we also used the duplication formula for the gamma function $\Gamma(l/2)$ ([16], Chapter V, §1.7, eq.(4)). The proposition follows recalling definition (64) of the constant C. □

Proposition C4. *For any positive integer l and even $d \geq 4$, it holds*

$$L_K = C^{-1} \; (P_{d/2}(-\triangle))^{-1} \; \mathbb{P}_A,$$

where C is the positive constant introduced in the text of Theorem C1, $\mathbb{P}_A$ is the orthogonal projector in $\mathcal{L}^2(S^{d-1})$ over the subspace of antisymmetric functions, and $P_{d/2}$ is the polynomial defined in equation (63).

Proof. The proposition can be proved comparing the eigensystems of the operators $\mathbb{P}_A$ and $\triangle$ and those of L_K given by Proposition C3. Due to rotational invariance, the eigenspaces of these operators are the homogenous harmonics $\mathcal{Y}_l(d)$.

The eigenvalues λ'_l of $\mathbb{P}_A$ are clearly determined by the degree of the harmonics, in fact

$$\lambda'_l = (1-(-1)^l)/2.$$

The eigenvalues λ''_l of the Laplacian $\triangle$ are (see for example [15] §15 Lemma 1)

$$\lambda''_l = -l(l+d-2).$$

Comparing these expressions with the eigenvalues of L_K, given by Proposition C3, it is clear that we are left with proving that, for odd l and even d,

$$2^d \Gamma(d/2)^2 B(l/2, d/2)^{-2} = P_{d/2}(l(l+d-2)). \qquad (70)$$

The last equation follows from Proposition C2 and observing that, we can write

$$B(l/2,d/2)^{-1} = \frac{\Gamma(\frac{l+d}{2})}{\Gamma(\frac{d}{2})\Gamma(\frac{l}{2})} = \frac{1}{2^{d/2}\Gamma(\frac{d}{2})} \frac{(l+d-2)!!}{(l-2)!!},$$

where we used the identity

$$\Gamma(k+1/2) = \sqrt{\pi}2^{-k}(2k-1)!!,$$

holding for positive integers k. □

Acknowledgments. The authors would like to acknowledge Ernesto De Vito and Adam Kalai for useful discussions and suggestions.

The authors are supported by the NSF grant 0325113. Andrea Caponnetto is also partially funded by the FIRB Project ASTAA and the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

[1] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov. Universal algorithms for learning theory. part i: piecewise constant functions. *Journal of Machine Learning Reasearch*, 6:1297–1321, 2005.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. Wadsworth International Group, 1984.

[3] V.I. Burenkov. *Sobolev spaces on domains*. B. G. Teubuer, Stuttgart-Leipzig, 1998. 312 ISBN 3-8154-2068-7.

[4] A. Caponnetto and Y. Yao. Adaptive learning algorithms via cross-validation. University of Chicago, *Preprint*, 2005.

[5] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.

[6] E. De Vito and A. Caponnetto. Risk bounds for regulaized least-squares algorithm with operator-valued kernels. Technical report, Massachusetts Institute of Technology, Cambridge, MA, May 2005. CBCL Paper #249/AI Memo #2005-015.

[7] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Series in Statistics, 2002.

[8] Minh Ha Quang. Personal communication.

[9] D. Haussler. Convolution kernels on discrete structures. Technical report, Dept. of CS, University of California at Santa Cruz, July 1999. UCSC-CRL-99-10.

[10] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[11] Adam Kalai. Efficient estimators for generalized additive models. *TTI-Chicago*, *preprint*, 2005.

[12] Adam Kalai. Learning monotonic linear functions. *Proceedings of 17th Annual Conference on Learning Theory, COLT 2004*.

[13] M. Kearns and Y. Mansour. On boosting ability of top-down decision tree learning algorithms. *Journal of Computer and System Sciences*, 58:109–128, 1999.

[14] Y. Mansour and D. McAllester. Boosting using branching programs. *Journal of Computer and System Sciences*, 64:103–112, 2002.

[15] Claus Müller. *Analysis of Spherical Symmetries in Euclidean Spaces*, volume 129 of *Applied Mathematical Sciences*. Springer, New York, 1998.

[16] N. Ja. Vilenkin. *Special Functions and the Theory of Group Representations*, volume 22 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, 1968.

Andrea Caponnetto, Department of Computer Science, University of Chicago, 1100 East 58th Street, Chicago, IL 60637. Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy

E-mail address: caponnet@uchicago.edu

Steve Smale, Toyota Technological Institute at Chicago, 1427 East 60th Street, Chicago,IL 60637

E-mail address: smale@math.berkeley.edu