
Kernel Partial Least Squares is Universally Consistent

Gilles Blanchard Nicole Krämer

Weierstrass Institute for Applied Analysis and Stochastics
Mohrenstr. 39, 10117 Berlin, Germany

Abstract

We prove the statistical consistency of kernel Partial Least Squares Regression applied to a bounded regression learning problem on a reproducing kernel Hilbert space. Partial Least Squares stands out of well-known classical approaches as e.g. Ridge Regression or Principal Components Regression, as it is not defined as the solution of a global cost minimization procedure over a fixed model nor is it a linear estimator. Instead, approximate solutions are constructed by projections onto a nested set of data-dependent subspaces. To prove consistency, we exploit the known fact that Partial Least Squares is equivalent to the conjugate gradient algorithm in combination with early stopping. The choice of the stopping rule (number of iterations) is a crucial point. We study two empirical stopping rules. The first one monitors the estimation error in each iteration step of Partial Least Squares, and the second one estimates the empirical complexity in terms of a condition number. Both stopping rules lead to universally consistent estimators provided the kernel is universal.

1 INTRODUCTION

Partial Least Squares (PLS) (Wold, 1975) is a supervised dimensionality reduction technique. It iteratively constructs an orthogonal set of m latent components from the predictor variables which have maximal covariance with the response variable. This low-dimensional representation of the data is then used for prediction by fitting a linear regression model to the

response and the latent components. The number m of latent components acts as a regularizer. In contrast to Principal Components Regression, the latent components are response-dependent. In combination with the kernel trick (Schölkopf et al., 1998), kernel PLS (KPLS) performs nonlinear dimensionality reduction and regression (Rosipal and Trejo, 2001).

While PLS has proven to be successful in a wide range of applications, theoretical studies of PLS – such as its consistency – are less widespread. This is perhaps due to the fact that in contrast to many standard methods (as e.g. Ridge Regression or Principal Components Regression), PLS is not defined as the solution of a global cost function nor is it a linear estimator in the sense that the fitted values depend linearly on the response variable. Instead, PLS minimizes the least squares criterion on a nested subset of data-dependent subspaces (i.e., the subspaces defined by the latent components). Therefore, results obtained for linear estimators are not straightforward to extend to PLS. Recent work (Naik and Tsai, 2000; Chun and Keles, 2010) study the model consistency of PLS. Their results assume that the target function depends on a finite known number ℓ of orthogonal latent components and that PLS is run at least for ℓ steps (without early stopping). In this configuration, Chun and Keles (2010) obtain inconsistency results in scenarios where the dimensionality can grow with the number of data. While the model considered here (KPLS) is different, this still underscores that the choice of the regularization (or early stopping) term m is important and that it has to be selected in a data-dependent manner.

Our focus in the present paper is the study of stopping rules that ensure universal consistency of KPLS in the infinite dimensional case. For the derivation of our results, we capitalize on the close connection of PLS and the conjugate gradient algorithm (Hestenes and Stiefel, 1952) for the solution of linear equations: The PLS solution with m latent components is equivalent to the conjugate algorithm applied to the set of normal equations in combination with early stopping after m iterations. We use this equivalence to define

Appearing in Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

the population version of KPLS. We then proceed in three steps: (i) We show that population KPLS converges to the true regression function. (ii) We bound the difference between empirical and population PLS, which is low as long as the number of iterations does not grow too fast. We ensure this via two different stopping criteria. The first one monitors the error in each iteration stop of PLS, and the second one estimates the empirical complexity in terms of a condition number. (iii) Combining the results from the two previous steps, our stopping rules lead to universally consistent estimators provided the kernel is universal. We emphasize that either stopping rule does not depend on any prior knowledge of the target function and only depends on observable quantities.

2 BACKGROUND

We study a regression problem based on a joint probability distribution $P(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$. The task is to estimate the true regression function

$$\bar{f}(x) = \mathbb{E}[Y|X = x] \quad (1)$$

based on a finite number n of observations $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$. As a general convention, population quantities defined from the perfect knowledge of the distribution P will be denoted with a bar, empirical quantities without. We assume that \bar{f} belongs to the space of P_X -square-integrable functions $\mathcal{L}_2(P_X)$, where P_X denotes the X -marginal distribution. The vector $\mathbf{y} \in \mathbb{R}^n$ represents the n centered response observations y_1, \dots, y_n .

As we consider kernel techniques to estimate the true regression function (1), we map the data to a reproducing kernel Hilbert space \mathcal{H}_k with bounded kernel k , via the canonical kernel map

$$\phi : \mathcal{X} \rightarrow \mathcal{H}_k, \quad x \mapsto \phi(x) = k(x, \cdot).$$

In the remainder, we make the following assumptions:

(B) boundedness of the data and kernel: $Y \in [-1, 1]$ almost surely and $\sup_{x \in \mathcal{X}} k(x, x) \leq 1$.

(U) universality of the kernel: for any distribution P_X on \mathcal{X} , \mathcal{H}_k is dense in $\mathcal{L}_2(P_X)$.

We remark that this is a weaker formulation of universality than the one defined by Steinwart (2001).

2.1 LEARNING AS AN INVERSE PROBLEM

We very briefly review the interpretation of kernel based regression as a statistical inverse problem, as introduced in De Vito et al. (2006). Let us denote the

inclusion operator of the kernel space into $\mathcal{L}_2(P_X)$ by

$$\bar{T} : \mathcal{H}_k \hookrightarrow \mathcal{L}_2(P_X) \quad g \mapsto g.$$

This operator maps a function to itself, but between two Hilbert spaces which differ with respect to their geometry – the inner product of \mathcal{H}_k being defined by the kernel function k , while the inner product of $\mathcal{L}_2(P_X)$ depends on the data generating distribution. The adjoint operator of \bar{T} is defined as usual as the unique operator satisfying the condition $\langle f, \bar{T}g \rangle_{\mathcal{L}_2(P_X)} = \langle \bar{T}^*f, g \rangle_{\mathcal{H}_k}$ for all $f \in \mathcal{L}_2(P_X), g \in \mathcal{H}_k$. It can be checked from this definition and the reproducing property of the kernel that \bar{T}^* coincides with the *kernel integral operator* from $\mathcal{L}_2(P_X)$ to \mathcal{H}_k ,

$$\bar{T}^*g = \int k(\cdot, x')g(x')dP(x') = \mathbb{E}[k(X, \cdot)g(X)].$$

Finally, the operator $\bar{S} = \bar{T}^*\bar{T} : \mathcal{H}_k \rightarrow \mathcal{H}_k$ is the *covariance operator* for the random variable $\phi(X)$:

$$Sg = \mathbb{E}[k(X, \cdot)g(X)] = \mathbb{E}[\phi(X) \langle \phi(X), g \rangle].$$

Learning in the kernel space \mathcal{H}_k can be cast (formally) as the inverse problem $\bar{T}\bar{g} = \bar{f}$, which yields (after right multiplication by \bar{T}^*) the so-called normal equation

$$\bar{S}\bar{g} = \bar{T}^*\bar{f}. \quad (2)$$

The above equation has a solution if and only if \bar{f} can be represented as a function of \mathcal{H}_k , i.e. $\bar{f} \in \bar{T}\mathcal{H}_k$; however, even if $\bar{f} \notin \bar{T}\mathcal{H}_k$, the above formal equation can be used as a motivation to use regularization algorithms coming from the inverse problems literature in order to find an *approximation* \bar{g} of \bar{f} belonging to the space \mathcal{H}_k . In a learning problem, neither the left-hand nor the right-hand side of (2) is known, and we only observe empirical quantities, which can be interpreted as "perturbed" versions of the population equations wherein P_X is replaced by its empirical counterpart $P_{X,n}$ and \bar{f} by \mathbf{y} . Note that the space $\mathcal{L}_2(P_{X,n})$ is isometric to \mathbb{R}^n with the usual Euclidean product, wherein a function $g \in \mathcal{L}_2(P_{X,n})$ is mapped to the n -vector $(g(x_1), \dots, g(x_n))$. The empirical integral operator $T^* : \mathcal{L}_2(P_{X,n}) \rightarrow \mathcal{H}_k$ is then given by

$$T^*g = \frac{1}{n} \sum_{i=1}^n g(x_i)k(x_i, \cdot). \quad (3)$$

The empirical covariance operator $S = T^*T$ is defined similarly, but on the input space \mathcal{H}_k . Note that if the Hilbert space \mathcal{H}_k is finite-dimensional, the operator S corresponds to left multiplication with the empirical covariance matrix. The perturbed, empirical version of the normal equation (2) is then defined as

$$Sg = T^*\mathbf{y}. \quad (4)$$

Again, if \mathcal{H}_k is finite-dimensional, the right-hand side corresponds to the covariance between predictors and response. In general, equation (4) is ill-posed, and regularization techniques are needed. Popular examples are Kernel Ridge Regression (which corresponds to Tikonov regularization in inverse problems) or ℓ_2 -Boosting (corresponding to Landweber iterations (Engl et al., 1996)).

2.2 PLS AND CONJUGATE GRADIENTS

PLS is generally described as a greedy iterative method that produces a sequence of latent components on which the data is projected. In contrast with PCA components, which maximize the variance, in PLS, the components are defined to have maximal covariance with the response \mathbf{y} . In particular, the latent components depend on the response.

It is however a known fact that the output of the m -th step of PLS is actually equivalent to a conjugate gradient (CG) algorithm applied to the normal equation (4), stopped early at step m (see e.g. Phatak and de Hoog, 2002 for a detailed overview). This has been established for traditional PLS, i.e. $\mathcal{X} = \mathbb{R}^d$ and the linear kernel is used, but for the kernel PLS algorithm introduced by Rosipal and Trejo (2001) the exact same analysis is valid as well. Here, for reasons of clarity with the remainder of our analysis we therefore directly present KPLS as a CG method.

For the self-adjoint operator S and for $T^*\mathbf{y}$, we define the associated Krylov space of order m as

$$\mathcal{K}_m(T^*\mathbf{y}, S) = \text{span}\{T^*\mathbf{y}, ST^*\mathbf{y}, \dots, S^{m-1}T^*\mathbf{y}\} \subset \mathcal{H}_k.$$

In other words, $\mathcal{K}_m(T^*\mathbf{y}, S)$ is the linear subspace of \mathcal{H}_k of all elements of the form $q(S)T^*\mathbf{y}$, where q ranges over the real polynomials of degree $m-1$. The m -th step of the CG method as applied to the normal equation (4) is simply defined (see e.g. Engl et al., 1996, chap. 7) as the element $g_m \in \mathcal{H}_k$ that minimizes the least squares criterion over the Krylov space, i.e.

$$g_m = \arg \min_{g \in \mathcal{K}_m(T^*\mathbf{y}, S)} \|\mathbf{y} - Tg\|^2. \quad (5)$$

Here, we recall that for any function $g \in \mathcal{H}_k$, the mapping Tg of g into $\mathcal{L}_2(P_{X,n})$ can be equivalently represented as the n -vector $(g(x_1), \dots, g(x_n))$. Observe that since the Krylov space depends itself on the data (and in particular on the response variable \mathbf{y}), (5) is not a linear estimator.

An extremely important property of CG is that the above optimization problem can be exactly computed by a simple iterative algorithm which only requires to use forward applications of the operator S , sums of

Algorithm 1 Empirical KPLS (in \mathcal{H}_k)

Initialize: $g_0 = 0; u_0 = T^*\mathbf{y}; d_0 = u_0$
for $m = 0, \dots, (m_{\max} - 1)$ **do**
 $\alpha_m = \|u_m\|^2 / \langle d_m, Sd_m \rangle$
 $g_{m+1} = g_m + \alpha_m d_m$ (update)
 $u_{m+1} = u_m - \alpha_m Sd_m$ (residuals)
 $\beta_m = \|u_{m+1}\|^2 / \|u_m\|^2$
 $d_{m+1} = u_{m+1} + \beta_m d_m$ (new basis vector)
end for
Return: approximate solution $g_{m_{\max}}$

elements in \mathcal{H}_k and scalar multiplications and divisions (algorithm 1).

In fact, the CG algorithm iteratively constructs a basis d_0, \dots, d_{m-1} of the Krylov space $\mathcal{K}_m(T^*\mathbf{y}, S)$. The sequence of $g_m \in \mathcal{H}_k$ is constructed in such a way that the residuals $u_m = T^*\mathbf{y} - Sg_m$ are pairwise orthogonal in \mathcal{H}_k , i.e. $\langle u_j, u_k \rangle = 0$ for $j \neq k$, while the constructed basis is S -orthogonal (or equivalently, uncorrelated), i.e. $\langle d_j, Sd_k \rangle = 0$ for $j \neq k$.

Note that the above algorithm is written entirely in \mathcal{H}_k , this form being convenient for the theoretical analysis to come. In practice, since all involved elements belong to $\text{span}\{(K(x_i, \cdot))_{1 \leq i \leq n}\}$, a weighted kernel expansion is used to represent these elements, and corresponding weight update equations using the kernel matrix can be derived (see Rosipal and Trejo, 2001).

3 POPULATION KPLS

Using the CG interpretation of KPLS, we can define its population version as follows:

Definition 1. Denote by $\bar{g}_m \in \mathcal{H}_k$ the output of algorithm 1 after m iterations, if we replace the empirical operator S and the vector $T^*\mathbf{y}$ by their population versions \bar{S} and T^*f , respectively. We define population KPLS with m components as $\bar{f}_m = T\bar{g}_m \in \mathcal{L}_2(P_X)$.

We emphasize again that $\bar{g}_m \in \mathcal{H}_k$ and $\bar{f}_m \in \mathcal{L}_2(P_X)$ are identical as functions from \mathcal{X} to \mathbb{R} , but seen as elements of Hilbert spaces with a different geometry (norm). The first step in our consistency proof is to show that population KPLS \bar{f}_m converges to \bar{f} (with respect to the $\mathcal{L}_2(P_X)$ norm) if m tends to ∞ . Note that even if $\bar{f} \notin T\mathcal{H}_k$, we can still show that $T\bar{g}_m$ converges to the projection of \bar{f} onto the closure of \mathcal{H}_k in $\mathcal{L}_2(P)$. If the kernel is universal (U), this projection is \bar{f} itself and this implies asymptotic consistency of the population version. We will assume for simplicity that

(I) the true regression function \bar{f} has infinitely many components in its decomposition over the eigenfunc-

tions of \bar{S} ,

which implies that the population version of the algorithm can theoretically be run indefinitely without exiting. If this condition is not satisfied the population algorithm stops after a finite number of steps κ , at which points it holds that $\bar{f}_\kappa = \bar{f}$ so that the rest of our analysis also holds in that case with only minor modifications.

Proposition 1. *The kernel operator of k is defined as $\bar{K} = \bar{T}\bar{T}^* : \mathcal{L}_2(P_X) \rightarrow \mathcal{L}_2(P_X)$. We denote by \mathcal{P} the orthogonal projection onto the closure of the range of \bar{K} in $\mathcal{L}_2(P)$. Then, recalling $\bar{f}_m = \bar{T}\bar{g}_m$ where \bar{g}_m is the output of the m -th iteration of the conjugate gradient algorithm applied to the population normal equation (2), it holds that $\bar{f}_m = \bar{K}q_m(\bar{K})\bar{f}$, where q_m is a polynomial of degree $\leq m-1$ fulfilling*

$$q_m = \arg \min_{\deg q \leq m-1} \|\mathcal{P}\bar{f} - \bar{K}q(\bar{K})\bar{f}\|_{\mathcal{L}_2(P_X)}^2.$$

Proof. The minimization property (5) when written in the population case yields

$$q_m = \arg \min_{\deg q \leq m-1} \|\bar{f} - \bar{T}q(\bar{S})\bar{T}^*\bar{f}\|^2.$$

Furthermore, for all polynomials q

$$\begin{aligned} \|\bar{f} - \bar{T}q(\bar{S})\bar{T}^*\bar{f}\|^2 &= \|\bar{f} - \bar{K}q(\bar{K})\bar{f}\|^2 \\ &= \|\mathcal{P}(\bar{f} - \bar{K}q(\bar{K})\bar{f})\|^2 \\ &\quad + \|(I - \mathcal{P})(\bar{f} - \bar{K}q(\bar{K})\bar{f})\|^2 \\ &= \|\mathcal{P}\bar{f} - \bar{K}q(\bar{K})\bar{f}\|^2 + \|(I - \mathcal{P})\bar{f}\|^2. \end{aligned}$$

As the second term above does not depend on the polynomial q , this yields the announced result. \square

This leads to the following convergence result.

Theorem 1. *Let us denote by \tilde{f}_m the projection of \bar{f} onto the first m principal components of the operator \bar{K} . We have*

$$\|\mathcal{P}\bar{f} - \tilde{f}_m\|_{\mathcal{L}_2(P_X)} \leq \|\mathcal{P}\bar{f} - \tilde{f}_m\|_{\mathcal{L}_2(P_X)}.$$

In particular,

$$\tilde{f}_m \xrightarrow{m \rightarrow \infty} \mathcal{P}\bar{f} \quad \text{in } \mathcal{L}_2(P_X).$$

This theorem is an extension of the finite-dimensional results by Phatak and de Hoog (2002).

Proof. We construct a sequence of polynomials \tilde{q}_m of degree $\leq m-1$ such that

$$\|\mathcal{P}\bar{f} - \bar{K}\tilde{q}_m(\bar{K})\bar{f}\| \leq \|\mathcal{P}\bar{f} - \tilde{f}_m\|$$

and then exploit the minimization property of Proposition 1. Let us consider the first m eigenvalues

$\lambda_1, \dots, \lambda_m$ of the operator \bar{K} with corresponding eigenfunctions ϕ_1, \dots, ϕ_m . Then, by definition

$$\tilde{f}_m = \sum_{i=1}^m \langle \bar{f}, \phi_i \rangle \phi_i, \quad \mathcal{P}\bar{f} = \sum_{i=1}^{\infty} \langle \bar{f}, \phi_i \rangle \phi_i,$$

The polynomial

$$\tilde{p}_m(\lambda) = \prod_{i=1}^m \frac{\lambda_i - \lambda}{\lambda_i} \quad (6)$$

fulfills $\tilde{p}_m(0) = 1$, hence it defines a polynomial \tilde{q}_m of degree $\leq m-1$ via $\tilde{p}_m(\lambda) = 1 - \lambda\tilde{q}_m(\lambda)$. As the zeroes of \tilde{p}_m are the first m eigenvalues of \bar{K} , the polynomial \tilde{q}_m has the convenient property that it "cancels out" the first m eigenfunctions, i.e.

$$\|\mathcal{P}\bar{f} - \bar{K}\tilde{q}_m(\bar{K})\bar{f}\|^2 = \sum_{i=m+1}^{\infty} \tilde{p}_m(\lambda_i)^2 \langle \bar{f}, \phi_i \rangle^2$$

By construction, $\tilde{p}_m(\lambda_i)^2 \leq 1$ for $i > m$, and hence

$$\|\mathcal{P}\bar{f} - \bar{K}\tilde{q}_m(\bar{K})\bar{f}\|^2 \leq \sum_{i=m+1}^{\infty} \langle \bar{f}, \phi_i \rangle^2 = \|\mathcal{P}\bar{f} - \tilde{f}_m\|^2.$$

As the principal components approximations \tilde{f}_m converge to $\mathcal{P}\bar{f}$, this concludes the proof. \square

This theorem shows in particular that the conjugate gradient method is less biased than Principal Components Analysis. This fact is known for linear PLS in the empirical case (De Jong, 1993; Phatak and de Hoog, 2002). By the same token, KPLS is less biased than ℓ_2 -Boosting, as the latter corresponds to the fixed polynomial $q_m(t) = \sum_{i=0}^{m-1} (1-t)^i$. However, empirical findings suggest that for KPLS, the decrease in bias is balanced by an increased complexity in terms of degrees of freedom (Kramer and Braun, 2007). The goal of the next section is to introduce a suitable control of this complexity.

4 CONSISTENT STOPPING RULES

4.1 ERROR MONITORING

We control the error between the population case and the empirical case by iteratively monitoring upper bounds on this error. Since this bound only involves known empirical quantities, we design a stopping criterion based on the bound. This then leads to a globally consistent procedure. The key ingredient of the stopping rule is to bound the differences for u, x, d (defined in algorithm 1) if we replace the empirical quantities S and $T^*\mathbf{y}$ by their population versions. Note that algorithm 1 involves products and quotients of the perturbed quantities. The error control based on these

expressions can hence be represented in terms of the following three functions.

Definition 2. For any positive reals $x > \delta_x \geq 0$ define

$$\zeta(x, \delta_x) = \frac{\delta_x}{x(x - \delta_x)}$$

and for any positive reals $(x, y, \delta_x, \delta_y)$ define

$$\begin{aligned} \xi(x, y, \delta_x, \delta_y) &= x\delta_y + y\delta_x + \delta_x\delta_y; \\ \xi'(x, y, \delta_x, \delta_y) &= x\delta_y + y\delta_x. \end{aligned}$$

The usefulness of these definitions is justified by the following standard lemma for bounding the approximation error of inverse and products, based only on the knowledge of the approximant:

Lemma 1. Let $\alpha, \bar{\alpha}$ be two invertible elements of a Banach algebra \mathcal{B} , with $\|\alpha - \bar{\alpha}\| \leq \delta$ and $\|\alpha^{-1}\|^{-1} > \delta$. Then

$$\|\alpha^{-1} - \bar{\alpha}^{-1}\| \leq \zeta(\|\alpha^{-1}\|^{-1}, \delta).$$

Let $\mathcal{B}_1, \mathcal{B}_2$ be two Banach spaces and assume an associative product operation exists from $\mathcal{B}_1 \times \mathcal{B}_2$ to a Banach space \mathcal{B}_3 , satisfying for any $(x_1, x_2) \in \mathcal{B}_1 \times \mathcal{B}_2$ the product compatibility condition $\|x_1 x_2\| \leq \|x_1\| \|x_2\|$. Let $\alpha, \bar{\alpha}$ in \mathcal{B}_1 and $\beta, \bar{\beta}$ in \mathcal{B}_2 , such that $\|\alpha - \bar{\alpha}\| \leq \delta_\alpha$ and $\|\beta - \bar{\beta}\| \leq \delta_\beta$. Then

$$\|\alpha\beta - \bar{\alpha}\bar{\beta}\| \leq \xi(\|\alpha\|, \|\beta\|, \delta_\alpha, \delta_\beta).$$

In the same situation as above, if it is known that $\|\bar{\alpha}\| \leq C$, then

$$\|\alpha\beta - \bar{\alpha}\bar{\beta}\| \leq \xi'(C, \|\beta\|, \delta_\alpha, \delta_\beta).$$

Furthermore, we can bound the deviation of the 'starting values' S and $T^* \mathbf{y}$:

Lemma 2. Set $\varepsilon_n = 4\sqrt{(\log n)/n}$. If the kernel is bounded **(B)**, with probability at least $1 - n^{-2}$,

$$\|T^* \mathbf{y} - \bar{T} \bar{f}\| \leq \varepsilon_n \quad \text{and} \quad \|S - \bar{S}\| \leq \varepsilon_n.$$

The second bound is well-known, see e.g. Shawe-Taylor and Cristianini (2003); Zwald and Blanchard (2006), and the first one is based on the same argument.

The error monitoring updates corresponding to algorithm 1 are displayed in algorithm 2. Note that the error monitoring initialization and update only depend on observable quantities.

Definition 3 (First stopping rule). Fix $0 < \gamma < \frac{1}{2}$ and run the KPLS algorithm 1 along with the error monitoring algorithm 2. Let $m_{(n)} + 1$ denote the first time where either the procedure exits, or $\delta_{m_{(n)}+1}^g > n^{-\gamma}$. Here, the subscript (n) indicates that the step is data-dependent. Output the estimate at the previous step, that is, the estimate $f^{(n)} = Tg_{m_{(n)}}$.

Algorithm 2 Error Control for Algorithm 1

Initialize: $\delta_0^g = 0; \varepsilon_n = 4\sqrt{(\log n)/n}; \delta_0^u = \varepsilon_n;$

$\delta_0^d = \varepsilon_n$

Initialize: $\varepsilon_{0,4} = \xi(\|u_0\|, \|u_0\|, \delta_0^u, \delta_0^u)$

for $m = 0, \dots, (m_{\max} - 1)$ **do**

$\varepsilon_{m,1} = \xi'(\|d_m\|, 1, \delta_m^d, \varepsilon_n)$

$\varepsilon_{m,2} = \xi(\|d_m\|, \|d_m\|, \delta_m^d, \varepsilon_{m,1})$

$\varepsilon_{m,3} = \zeta(\langle d_m, Sd_m \rangle, \varepsilon_{m,2})$ (if defined, else exit)

$\delta_m^\alpha = \xi(\|u_m\|^2, \langle d_m, Sd_m \rangle^{-1}, \varepsilon_{m,4}, \varepsilon_{m,3})$

$\delta_{m+1}^g = \delta_m^g + \xi(\alpha_m, \|d_m\|, \delta_m^\alpha, \delta_m^d)$

$\delta_{m+1}^u = \delta_m^u + \xi(\alpha_m, \|d_m\|, \delta_m^\alpha, \varepsilon_{m,1})$

$\varepsilon_{m,5} = \zeta(\|u_m\|^2, \varepsilon_{m,4})$ (if defined, else exit)

$\varepsilon_{m+1,4} = \xi(\|u_{m+1}\|, \|u_{m+1}\|, \delta_{m+1}^u, \delta_{m+1}^u)$

$\delta_m^\beta = \xi(\|u_{m+1}\|^2, \|u_m^{-1}\|^2, \varepsilon_{m+1,4}, \varepsilon_{m,5})$

$\delta_{m+1}^d = \delta_m^d + \xi(\beta_m, \|d_m\|, \delta_m^\beta, \delta_m^d)$

end for

The next theorem states that this stopping rule leads to a universally consistent learning rule for bounded regression.

Theorem 2. Assume that the kernel is bounded **(B)** and universal **(U)**. Denote by $f^{(n)}$ the output of the KPLS run on n independent observations while using the above stopping rule. Then almost surely

$$\lim_{n \rightarrow \infty} \left\| \bar{f} - f^{(n)} \right\|_{\mathcal{L}_2(P_X)} = 0.$$

Proof. By construction $f^{(n)} = Tg_{m_{(n)}}^{(n)}$. We have

$$\left\| \bar{f} - f^{(n)} \right\| \leq \left\| \bar{f} - \bar{f}_{m_{(n)}} \right\| + \left\| \bar{f}_{m_{(n)}} - Tg_{m_{(n)}}^{(n)} \right\|.$$

We proceed in two steps. First, we establish that the second term is bounded by $n^{-\gamma}$. Second, we prove that the random variable $m_{(n)} \rightarrow \infty$ almost surely for $n \rightarrow \infty$, which ensures that the first term goes to zero (using Theorem 1).

First Step: We have

$$\begin{aligned} \left\| \bar{f}_{m_{(n)}} - Tg_{m_{(n)}}^{(n)} \right\|_{\mathcal{L}_2(P_X)} &= \left\| T(\bar{g}_m - g_m) \right\|_{\mathcal{L}_2(P_X)} \\ &\leq \left\| \bar{g}_m - g_m \right\|_\infty \\ &\leq \left\| \bar{g}_m - g_m \right\|_{\mathcal{H}_k}. \end{aligned}$$

(We drop the superscript (n) for m to lighten notation.) Now, we prove that the construction of the error monitoring iterates ensures that $\|\bar{g}_m - g_m\| \leq \delta_m^g$ for any m before stopping, with probability at least $1 - 2n^{-2}$. For $m = 0$, this follows immediately from Lemma 2. For a general m , it is then straightforward to show that $\varepsilon_{m,1}$ controls the error for Sd_m (using $\|S\| \leq 1$); $\varepsilon_{m,2}$ controls the error for $\langle \bar{d}_k, S\bar{d}_k \rangle$ (the Cauchy-Schwartz inequality ensuring the compatibility condition of norm and product in Lemma 1);

$\varepsilon_{m,3}$ controls the error for the inverse of the latter; $\delta_m^\alpha, \delta_m^g, \delta_m^u, \delta_m^\beta, \delta_m^d$ control the errors for the respective superscript quantities; $\varepsilon_{m,4}$ controls the error for $\|\bar{u}_m\|^2$ and $\varepsilon_{m,5}$ for the inverse of the latter. In particular, with probability at least $1 - 2n^{-2}$ we have $\|g_m - \bar{g}_m\| \leq \delta_m^g$, so that $\|g_{m(n)} - \bar{g}_{m(n)}\| \leq n^{-\gamma}$ by definition of the stopping rule. Since the probabilities that this inequalities are violated are summable, by the Borel-Cantelli lemma, almost surely this inequality is true from a certain rank n_0 on.

Second Step: This is in essence a minutely detailed continuity argument. Due to space limitations, we only sketch the proof. We consider a fixed iteration m and prove that this iteration is reached without exiting and that $\delta_m^{g(n)} < C(m)\varepsilon_n$ almost surely from a certain rank n_0 on (here the superscript (n) once again recalls the dependence on the number of data). The constant $C(m)$ is deterministic but can depend on the generating distribution. This obviously ensures $m(n) \rightarrow \infty$ almost surely. We prove by induction that this property is true for all error controlling quantities $\varepsilon_{*,m}$ and δ_m^* appearing in the error monitoring iterates. Obviously, from the initialization this is true for $\delta_0^g, \delta_0^u, \delta_0^d$ and ε_n . In a nutshell, the induction step then follows from the fact that the error monitoring functions ζ, ξ, ξ' are locally Lipschitz. \square

4.2 EMPIRICAL COMPLEXITY

We now propose an alternate stopping rule directly based on the characterizing property (5) of KPLS/CG. This approach leads to a more explicit stopping rule. Let us define the operator $R_m : \mathbb{R}^m \rightarrow \mathcal{H}$:

$$v = (v_1, \dots, v_m) \mapsto R_m v = \sum_{i=1}^m v_i S^{i-1} T^* \mathbf{y}. \quad (7)$$

Then, we can rewrite the m -th iterate of KPLS/CG as $g_m = R_m w_m$, and (5) becomes

$$w_m = \arg \min_{w \in \mathbb{R}^m} \|\mathbf{y} - T R_m w\|^2.$$

From this, it can be deduced by standard arguments that

$$g_m = R_m M_m^{-1} R_m^* T^* \mathbf{y}, \quad M_m = R_m^* S R_m. \quad (8)$$

The random $m \times m$ matrix M_m has (i, j) entry

$$(M_m)_{ij} = \mathbf{y}^\top T S^{i-1} S S^{j-1} T^* \mathbf{y} = \mathbf{y}^\top K^{(i+j)} \mathbf{y},$$

where $K = T T^*$ can be identified with the kernel Gram matrix, i.e. $K_{k\ell} = k(x_k, x_\ell)$. Similarly, denote by M'_m the matrix with entry (i, j) equal to $\mathbf{y}^\top K^{(i+j-1)} \mathbf{y}$.

Definition 4 (Second stopping rule). Fix $\frac{1}{2} > \nu > 0$ and let $m'_{(n)} + 1$ denote the first time where $m (\max(\|M'_m\|, m^{-1}) \|M_m^{-1}\|)^2 \geq n^\nu$. (If M_m is singular, we set $\|M_m^{-1}\| = \infty$.) Output the KPLS estimate at step $m'_{(n)}$.

Note that the stopping criterion only depends on empirical quantities. Furthermore, we can interpret the stopping criterion as an empirical complexity control of the CG algorithm at step m . Essentially, it is the dimensionality (or iteration number) m times the square of the “pseudo-condition number” $\|M'_m\| \|M_m^{-1}\|$.

Theorem 3. Assume that the kernel is bounded (B) and universal (U). Denote by $f^{(n)}$ the output of the KPLS algorithm run on n independent observations while using the above stopping rule. Then almost surely

$$\lim_{n \rightarrow \infty} \left\| \bar{f} - f^{(n)} \right\|_{\mathcal{L}_2(P_X)} = 0.$$

Proof. We prove this result by exploiting the explicit formula (8) for KPLS. Note that the formula is also true for the population version \bar{x}_m if we replace all empirical quantities by their population counterparts. Hence, we need to control the deviation of the different terms appearing in the above formula (8). This boils down to perturbation analysis and is closely related to techniques employed by Chun and Keles (2010) in a finite dimensional context. The following lemma summarizes the deviation control (the proof can be found in the appendix).

Lemma 3. Put $\varepsilon_n = 4\sqrt{(\log n)/n}$. There is an absolute numerical constant c such that, with probability at least $1 - n^{-2}$, whenever $cm\varepsilon_n (\max(\|M'_m\|, m^{-1}) \|M_m^{-1}\|)^2 \leq 1$, then

$$\|g_m - \bar{g}_m\| \leq cm\varepsilon_n (\max(\|M'_m\|, m^{-1}) \|M_m^{-1}\|)^2.$$

Now, let c be the universal constant appearing in Lemma 3. At the stopping step $m = m'_{(n)}$, by construction,

$$cm\varepsilon_n (\max(\|M'_m\|, m^{-1}) \|M_m^{-1}\|)^2 \leq 4c \frac{n^{\nu-\frac{1}{2}}}{\sqrt{\log n}}.$$

Choosing some positive $\gamma < \frac{1}{2} - \nu$, for n large enough the right hand side is bounded by $n^{-\gamma}$. By Lemma 3 and a reasoning similar to that of the proof of Theorem 2, this implies the bound on the estimation error with respect to the population version (valid with probability at least $1 - n^{-2}$):

$$\left\| \bar{f}_{m'_{(n)}} - f_{m'_{(n)}} \right\| \leq n^{-\gamma}.$$

By the Borel-Cantelli lemma, this inequality is therefore almost surely satisfied for big enough n .

To conclude, we establish that $m'_{(n)} \rightarrow \infty$ almost surely as $n \rightarrow \infty$. It suffices to show that for any fixed m , for n larger than a certain $n_0(m)$ on, $m(\max(\|M'_m\|, m^{-1}) \|M_m^{-1}\|)^2 \leq n^\nu$, implying that almost surely $m'_{(n)} \geq m$ for n large enough. For this we simply show that the LHS of this inequality converges a.s. to a fixed number. From (9) in the proof of Lemma 3, and the straightforward inequality $\|M'\| \leq m$, we see that for a fixed iteration m , $\|M_m - \bar{M}_m\| \rightarrow 0$ almost surely as $n \rightarrow \infty$. The matrix \bar{M}_m is non-singular as we assume that the population version of the algorithm does not exit. This implies that M_m is almost surely non-singular for n big enough with $\|M_m^{-1} - \bar{M}_m^{-1}\| \rightarrow 0$, and therefore almost surely $\|M_m^{-1}\|$ converges to $\|\bar{M}_m^{-1}\|$, a fixed number. \square

5 CONCLUSION

We proposed two stopping rules for the number m of KPLS iterations that lead to universally consistent estimates. Both are based on the facts that (a) population KPLS is defined in terms of the covariance operator \bar{S} and the cross-covariance $\bar{T}^* \bar{f}$, and that (b) we can control the difference between population KPLS and empirical KPLS solutions based on the discrepancy of these covariances to their empirical counterparts S and $T^* \mathbf{y}$. Naturally, as for a large majority of learning algorithms, cross-validation remains the standard approach for model selection in practice. We want to stress that the motivation of the present work is mainly theoretical: universally consistent rules provide a strong learning theoretical foundation for KPLS.

Our approach makes heavy use of the equivalence of PLS to the conjugate gradient algorithm applied to the normal equations in combination with early stopping. This framework also connects KPLS to statistical inverse problems. In this context, KPLS stands out of previously well-studied classes of methods. In particular, as KPLS is not linear in \mathbf{y} , it contrasts the class of *linear* spectral methods for statistical inverse problems (e.g. Bissantz et al., 2007; Lo Gerfo et al., 2008). This class considers estimates of the form $g_\lambda = F_\lambda(S)T^* \mathbf{y}$ to regularize the ill-posed problem (4). Here, F_λ is a fixed family of "filter functions". Examples include Ridge Regression (also known as Thikonov regularization), Principal Components Regression (also known as spectral cut-off), and ℓ_2 -Boosting (which corresponds to Landweber iteration). In this paper, we explained that for KPLS with m components, the filter function F_λ is a polynomial of degree $\leq m - 1$, but that – unlike the class of linear spectral methods – this polynomial strongly depends on the response. For this reason, general results for linear spectral meth-

ods do not apply to KPLS, and additional techniques are needed. Furthermore, it is known in the deterministic setting that it is not possible to define a priori (i.e. data-independent) stopping rules that ensure consistency of CG (Engl et al., 1996, p.182). Similarly, the stopping rules proposed here strongly depend on the data, in contrast with other algorithms (e.g. nearest-neighbor) where a data-independent scheme $m(n)$ "growing slowly enough" is sufficient to ensure consistency.

For linear spectral methods, optimal convergence rates of the resulting estimators have been established in the recent years (Caponnetto and De Vito, 2007; Bauer et al., 2007; Caponnetto, 2006). Obviously, beyond consistency the focus of future theoretical effort on KPLS will be to establish that this algorithm can attain these optimal rates as well.

Acknowledgement

This work is funded in part by the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886.

References

- Bauer, F., Pereverzev, S., and Rosasco, L. (2007). On regularization algorithms in learning theory. *Journal of Complexity*, 23:52–72.
- Bissantz, N., Hohage, T., Munk, A., and Ruymgaart, F. (2007). Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636.
- Caponnetto, A. (2006). Optimal rates for regularization operators in learning theory. Technical Report CBCL Paper 264/ CSAIL-TR 2006-062, Massachusetts Institute of Technology.
- Caponnetto, A. and De Vito, E. (2007). Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.
- Chun, H. and Keles, S. (2010). Sparse partial least squares for simultaneous dimension reduction and variable selection. *JRSS B*, 72(1):3–25.
- De Jong, S. (1993). PLS fits closer than PCR. *Journal of chemometrics*, 7(6):551–557.
- De Vito, E., Rosasco, L., Caponnetto, A., De Giovannini, U., and Odone, F. (2006). Learning from Examples as an Inverse Problem. *Journal of Machine Learning Research*, 6(1):883.
- Engl, H., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems*. Kluwer Academic Publishers.

- Hestenes, M. and Stiefel, E. (1952). Methods for Conjugate Gradients for Solving Linear Systems. *Journal of Research of the National Bureau of Standards*, 49:409–436.
- Krämer, N. and Braun, M. (2007). Kernelizing PLS, Degrees of Freedom, and Efficient Model Selection. In *24th ICML*, pages 441 – 448.
- Lo Gerfo, L., Rosasco, L., Odone, F. and De Vito, E., and Verri, A. (2008). Spectral algorithms for supervised learning. *Neural Computation*, 20:1873–1897.
- Naik, P. and Tsai, C. (2000). Partial least squares estimator for single-index models. *Journal of the Royal Statistical Society: Series B*, 62(4):763–771.
- Phatak, A. and de Hoog, F. (2002). Exploiting the connection between PLS, Lanczos, and conjugate gradients: alternative proofs of some properties of PLS. *Journal of Chemometrics*, 16:361–367.
- Rosipal, R. and Trejo, L. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Spaces. *JMLR*, 2:97–123.
- Schölkopf, B., Smola, A. J., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319.
- Shawe-Taylor, J. and Cristianini, N. (2003). Estimating the moments of a random vector with applications. *Proceedings of GRETSI 2003 Conference*, pages pp. 47-52.
- Steinwart, I. (2001). On the Influence of the Kernel on the Consistency of Support Vector Machines. *Journal of Machine Learning Research*, 2:67–93.
- Wold, H. (1975). Path models with Latent Variables: The NIPALS Approach. In et al., H. B., editor, *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, pages 307–357. Academic Press.
- Zwald, L. and Blanchard, G. (2006). On the convergence of eigenspaces in kernel principal component analysis. In Weiss, Y., Bottou, L., and Platt, J., editors, *NIPS 2005*, pages 1649–1656. MIT Press.

A PROOF OF LEMMA 3

The letter c denotes an absolute numerical constant whose value can possibly be different from line to line.

We define \bar{R}_m as the population analogue of R_m , by replacing S and $T^*\mathbf{y}$ in (7) by \bar{S} and $\bar{T}^*\bar{f}$. The population version of KPLS is then

$$\bar{x}_m = \bar{R}_m \bar{M}_m^{-1} \bar{R}_m^* \bar{T}^* \bar{f}$$

where $\bar{M}_m = \bar{R}_m^* \bar{S} \bar{R}_m$. Since the index m of the iteration is now fixed, we omit the subscript m in

R and M . Set $\Delta = \max(\|M'\|, m^{-1})$. We have $M' = RR^*$, so that $\|M'\| = \|R\|^2$. Observe that $M' - M = R_m(Id - S)R_m^*$ is a positive-semidefinite matrix since $\|S\| \leq 1$; hence $\|M'\| \geq \|M\| \geq \|M^{-1}\|^{-1}$, and the assumption that $C\varepsilon_n m (\Delta \|M^{-1}\|)^2 \leq 1$ with $C \geq 1$ implies in particular that $m\varepsilon_n \leq 1$.

We first control the difference $\rho = \|R - \bar{R}\|$.

$$\begin{aligned} \rho &= \sup_{v:\|v\|=1} \left\| \sum_{i=1}^m v_i (S^{i-1} T^* \mathbf{y} - \bar{S}^{i-1} \bar{T}^* \bar{f}) \right\| \\ &\leq \sup_{v:\|v\|=1} \left(\left\| \sum_{i=1}^m v_i (S^{i-1} - \bar{S}^{i-1}) T^* \mathbf{y} \right\| \right) \\ &\quad + \sup_{v:\|v\|=1} \left\| \sum_{i=1}^m v_i \bar{S}^{i-1} (T^* \mathbf{y} - \bar{T}^* \bar{f}) \right\|. \end{aligned}$$

Using lemma 2, the second term on the right-hand side can be bounded by $\sum_{i=1}^m |v_i| \varepsilon_n \leq \|v\| \sqrt{m} \varepsilon_n$. For the first term, one can rewrite

$$\begin{aligned} &\sup_{v:\|v\|=1} \left\| \sum_{i=1}^m v_i (S^{i-1} - \bar{S}^{i-1}) T^* \mathbf{y} \right\| \\ &= \sup_{v:\|v\|=1} \left\| \sum_{i,j=1}^{m-1} \mathbf{1}\{j \leq i\} v_i \bar{S}^{j-1} (S - \bar{S}) \bar{S}^{i-j} T^* \mathbf{y} \right\| \\ &= \sup_{v:\|v\|=1} \left\| \sum_{j=1}^{m-1} \bar{S}^{j-1} (S - \bar{S}) \sum_{i=j}^{m-1} v_i \bar{S}^{i-j} T^* \mathbf{y} \right\| \\ &\leq \sum_{j=1}^{m-1} \varepsilon_n \sup_{v:\|v\|=1} \left\| \sum_{i=1}^{m-j} v_i \bar{S}^{i-1} T^* \mathbf{y} \right\| \leq m\varepsilon_n \|R\|. \end{aligned}$$

This implies

$$\|R - \bar{R}\| \leq \sqrt{m} \varepsilon_n + m\varepsilon_n \|R\| \leq 2m\varepsilon_n \Delta^{\frac{1}{2}}.$$

We now use repeatedly Lemma 1 to derive the final estimate from the above one using product and inverse operations. We omit some tedious details in the computations below.

Recalling $M = R^* S R$ and using $m\varepsilon_n \leq 1$, we deduce

$$\|M - \bar{M}\| \leq c(m\varepsilon_n \Delta + m^2 \varepsilon_n^2 \Delta) \leq cm\varepsilon_n \Delta. \quad (9)$$

We then have

$$\|M^{-1} - \bar{M}^{-1}\| \leq \zeta(\|M^{-1}\|^{-1}, cm\varepsilon_n \Delta) \leq cm\varepsilon_n \Delta \|M^{-1}\|^2.$$

Here, we assume that $cm\varepsilon_n \Delta \leq \|M^{-1}\|^{-1}/2$, which is implied by the assumption $C\varepsilon_n m \Delta^2 \|M^{-1}\|^2 \leq 1$ if C is chosen big enough. Finally, we get

$$\|RM^{-1}R^* - \bar{R}\bar{M}^{-1}\bar{R}^*\| \leq cm\varepsilon_n \Delta^2 \|M^{-1}\|^2.$$