Optimal learning rates for Kernel Conjugate Gradient regression

Gilles Blanchard
Mathematics Institute, University of Potsdam
Am neuen Palais 10, 14469 Potsdam
blanchard@math.uni-potsdam.de

Nicole Krämer
Weierstrass Institute
Mohrenstr. 39, 10117 Berlin, Germany
nicole.kraemer@wias-berlin.de

Abstract

We prove rates of convergence in the statistical sense for kernel-based least squares regression using a conjugate gradient algorithm, where regularization against overfitting is obtained by early stopping. This method is directly related to Kernel Partial Least Squares, a regression method that combines supervised dimensionality reduction with least squares projection. The rates depend on two key quantities: first, on the regularity of the target regression function and second, on the effective dimensionality of the data mapped into the kernel space. Lower bounds on attainable rates depending on these two quantities were established in earlier literature, and we obtain upper bounds for the considered method that match these lower bounds (up to a log factor) if the true regression function belongs to the reproducing kernel Hilbert space. If this assumption is not fulfilled, we obtain similar convergence rates provided additional unlabeled data are available. The order of the learning rates match state-of-the-art results that were recently obtained for least squares support vector machines and for linear regularization operators.

1 Introduction

The contribution of this paper is the learning theoretical analysis of kernel-based least squares regression in combination with conjugate gradient techniques. The goal is to estimate a regression function f^* based on random noisy observations. We have an i.i.d. sample of n observations $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ from an unknown distribution $P(X, Y)$ that follows the model

$$Y \quad = \quad f^*(X) + \varepsilon \,,$$

where ε is a noise variable whose distribution can possibly depend on X, but satisfies $\mathbb{E}\left[\varepsilon | X\right] = 0$. We assume that the true regression function f^* belongs to the space $\mathcal{L}_2(P_X)$ of square-integrable functions. Following the kernelization principle, we implicitly map the data into a reproducing kernel Hilbert space $\mathcal{H}$ with a kernel k. We denote by $K_n = \frac{1}{n}(k(X_i, X_j)) \in \mathbb{R}^{n \times n}$ the normalized kernel matrix and by $\Upsilon = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ the n-vector of response observations. The task is to find coefficients α such that the function defined by the normalized kernel expansion

$$f_\alpha(X) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \alpha_i k(X_i, X)$$

is an adequate estimator of the true regression function f^*. The closeness of the estimator f_α to the target f^* is measured via the $\mathcal{L}_2(P_X)$ distance,

$$\|f_\alpha - f^*\|_2^2 = \mathbb{E}_{X \sim P_X}\left[(f_\alpha(X) - f^*(X))^2\right] = \mathbb{E}_{XY}\left[(f_\alpha(X) - Y)^2\right] - \mathbb{E}_{XY}\left[(f^*(X) - Y)^2\right] \,,$$

The last equality recalls that this criterion is the same as the excess generalization error for the squared error loss $\ell(f, x, y) = (f(x) - y)^2$.

In empirical risk minimization, we use the training data empirical distribution as a proxy for the generating distribution, and minimize the *training* squared error. This gives rise to the linear equation

$$K_n \alpha = \Upsilon \qquad \text{with } \alpha \in \mathbb{R}^n \,. \tag{1}$$

Assuming K_n invertible, the solution of the above equation is given by $\alpha = K_n^{-1}\Upsilon$, which yields a function in $\mathcal{H}$ interpolating perfectly the training data but having poor generalization error. It is well-known that to avoid overfitting, some form of regularization is needed. There is a considerable variety of possible approaches (see e.g. [10] for an overview). Perhaps the most well-known one is

$$\alpha = (K_n + \lambda I)^{-1}\Upsilon, \tag{2}$$

known alternatively as kernel ridge regression, Tikhonov's regularization, least squares support vector machine, or MAP Gaussian process regression. A powerful generalization of this is to consider

$$\alpha = F_\lambda(K_n)\Upsilon, \tag{3}$$

where $F_\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ is a fixed function depending on a parameter λ. The notation $F_\lambda(K_n)$ is to be interpreted as F_λ applied to each eigenvalue of K_n in its eigen decomposition. Intuitively, F_λ should be a "regularized" version of the inverse function $F(x) = x^{-1}$. This type of regularization, which we refer to as linear regularization methods, is directly inspired from the theory of inverse problems. Popular examples include as particular cases kernel ridge regression, principal components regression and L_2-boosting. Their application in a learning context has been studied extensively [1, 2, 5, 6, 12]. Results obtained in this framework will serve as a comparison yardstick in the sequel.

In this paper, we study conjugate gradient (CG) techniques in combination with early stopping for the regularization of the kernel based learning problem (1). The principle of CG techniques is to restrict the learning problem onto a nested set of data-dependent subspaces, the so-called Krylov subspaces, defined as

$$\mathcal{K}_m(\Upsilon, K_n) = \text{span}\left\{\Upsilon, K_n\Upsilon, \ldots, K_n^{m-1}\Upsilon\right\} \,. \tag{4}$$

Denote by $\langle ., . \rangle$ the usual euclidean scalar product on $\mathbb{R}^n$ rescaled by the factor n^{-1}. We define the K_n-norm as $\|\alpha\|_{K_n}^2 := \langle \alpha, \alpha \rangle_{K_n} := \langle \alpha, K_n\alpha \rangle$. The CG solution after m iterations is formally defined as

$$\alpha_m = \arg \min_{\alpha \in \mathcal{K}_m(\Upsilon, K_n)} \|\Upsilon - K_n\alpha\|_{K_n} \,; \tag{5}$$

and the number m of CG iterations is the model parameter. To simplify notation we define $f_m := f_{\alpha_m}$. In the learning context considered here, regularization corresponds to early stopping. Conjugate gradients have the appealing property that the optimization criterion (5) can be computed by a simple iterative algorithm that constructs basis vectors $d_1, \ldots, d_m$ of $\mathcal{K}_m(\Upsilon, K_n)$ by using only *forward multiplication* of vectors by the matrix K_n. Algorithm 1 displays the computation of the CG kernel coefficients α_m defined by (5).

Algorithm 1 Kernel Conjugate Gradient regression

Input kernel matrix K_n, response vector Υ, maximum number of iterations m
Initialization: $\alpha_0 = \mathbf{0}_n; r_1 = \Upsilon; d_1 = \Upsilon; t_1 = K_n\Upsilon$
for $i = 1, \ldots, m$ **do**
  $t_i = t_i / \|t_i\|_{K_n}$; $d_i = d_i / \|t_i\|_{K_n}$ (normalization of the basis, resp. update vector)
  $\gamma_i = \langle \Upsilon, t_i \rangle_{K_n}$ (proj. of Υ on basis vector)
  $\alpha_i = \alpha_{i-1} + \gamma_i d_i$ (update)
  $r_{i+1} = r_i - \gamma_i t_i$ (residuals)
  $d_{i+1} = r_{i+1} - d_i \langle t_i, K_n r_{i+1} \rangle_{K_n}$; $t_{i+1} = K_n d_{i+1}$ (new update, resp. basis vector)
end for
Return: CG kernel coefficients α_m, CG function $f_m = \sum_{i=1}^n \alpha_{i,m} k(X_i, \cdot)$

The CG approach is also inspired by the theory of inverse problems, but it is not covered by the framework of linear operators defined in (3): As we restrict the learning problem onto the Krylov space $\mathcal{K}_m(\Upsilon, K_n)$, the CG coefficients α_m are of the form $\alpha_m = q_m(K_n)\Upsilon$ with q_m a polynomial of degree $\leq m - 1$. However, the polynomial q_m is not fixed but depends on Υ as well, making the CG method nonlinear in the sense that the coefficients α_m depend on Υ in a nonlinear fashion.

We remark that in machine learning, conjugate gradient techniques are often used as fast solvers for operator equations, e.g. to obtain the solution for the regularized equation (2). We stress that in this paper, we study conjugate gradients as a *regularization approach* for kernel based learning, where the regularity is ensured via early stopping. This approach is not new. As mentioned in the abstract, the algorithm that we study is closely related to Kernel Partial Least Squares [18]. The latter method also restricts the learning problem onto the Krylov subspace $\mathcal{K}_m(\Upsilon, K_n)$, but it minimizes the euclidean distance $\|\Upsilon - K_n\alpha\|$ instead of the distance $\|\Upsilon - K_n\alpha\|_{K_n}$ defined above[1]. Kernel Partial Least Squares has shown competitive performance in benchmark experiences (see e.g [18, 19]). Moreover, a similar conjugate gradient approach for non-definite kernels has been proposed and empirically evaluated by Ong et al [17]. The focus of the current paper is therefore not to stress the usefulness of CG methods in practical applications (and we refer to the above mentioned references) but to examine its theoretical convergence properties. In particular, we establish the existence of early stopping rules that lead to optimal convergence rates. We summarize our main results in the next section.

2 Main results

For the presentation of our convergence results, we require suitable assumptions on the learning problem. We first assume that the kernel space $\mathcal{H}$ is separable and that the kernel function is measurable. (This assumption is satisfied for all practical situations that we know of.) Furthermore, for all results, we make the (relatively standard) assumption that the kernel is *bounded*: $k(x,x) \leq \kappa$ for all $x \in \mathcal{X}$. We consider – depending on the result – one of the following assumptions on the *noise*:

(Bounded) (Bounded Y): $|Y| \leq M$ almost surely.

(Bernstein) (Bernstein condition): $\mathbb{E}\left[\varepsilon^p | X\right] \leq (1/2)p!M^p$ almost surely, for all integers $p \geq 2$.

The second assumption is weaker than the first. In particular, the first assumption implies that not only the noise, but also the target function f^* is bounded in supremum norm, while the second assumption does not put any additional restriction on the target function.

The *regularity* of the target function f^* is measured in terms of a *source condition* as follows. The kernel integral operator is given by

$$K : \mathcal{L}_2(P_X) \to \mathcal{L}_2(P_X),\ g \mapsto \int k(.,x)g(x)dP(x)\,.$$

The **source condition** for the parameters $r > 0$ and $\rho > 0$ is defined by:

$$\textbf{SC}(r,\rho) : f^* = K^r u \qquad \text{with} \qquad \|u\| \leq \kappa^{-r}\rho.$$

It is a known fact that if $r \geq 1/2$, then f^* coincides almost surely with a function belonging to $\mathcal{H}_k$. We refer to $r \geq 1/2$ as the "inner case" and to $r < 1/2$ as the "outer case".

The regularity of the kernel operator K with respect to the marginal distribution P_X is measured in terms of the so-called **effective dimensionality** condition, defined by the two parameters $s \in (0,1)$, $D \geq 0$ and the condition

$$\textbf{ED}(s,D) : \operatorname{tr}(K(K+\lambda I)^{-1}) \leq D^2(\kappa^{-1}\lambda)^{-s} \text{ for all } \lambda \in (0,1].$$

This notion was first introduced in [22] in a learning context, along with a number of fundamental analysis tools which we rely on and have been used in the rest of the related literature cited here. It is known that the best attainable rates of convergence, as a function of the number of examples n, are determined by the parameters r and s in the above conditions: It was shown in [6] that the minimax learning rate given these two parameters is lower bounded by $\mathcal{O}(n^{-2r/(2r+s)})$.

We now expose our main results in different situations. In all the cases considered, the early stopping rule takes the form of a so-called **discrepancy stopping rule:** For some sequence of thresholds $\Lambda_m > 0$ to be specified (and possibly depending on the data), define the (data-dependent) stopping iteration $\widehat{m}$ as the first iteration m for which

$$\|\Upsilon - K_n\alpha_m\|_{K_n} < \Lambda_m\,. \tag{6}$$

[1]This is generalized to a CG-l algorithm ($l \in \mathbb{N}_{\geq 0}$) by replacing the K_n-norm in (5) with the norm defined by K_n^l. Corresponding fast iterative algorithms to compute the solution exist for all l (see e.g. [11]).

Only in the first result below, the threshold Λ_m actually depends on the iteration m and on the data. It is not difficult to prove from (4) and (5) that $\|\Upsilon - K_n\alpha_n\|_{K_n} = 0$, so that the above type of stopping rule always has $\widehat{m} \leq n$.

2.1 Inner case without knowledge on effective dimension

The inner case corresponds to $r \geq 1/2$, i.e. the target function f^* lies in $\mathcal{H}$ almost surely. For some constants $\tau > 1$ and $1 > \gamma > 0$, we consider the discrepancy stopping rule with the threshold sequence

$$\Lambda_m = 4\tau\sqrt{\frac{\kappa\log(2\gamma^{-1})}{n}}\left(\sqrt{\kappa}\,\|\alpha_m\|_{K_n} + M\sqrt{\log(2\gamma^{-1})}\right). \tag{7}$$

For technical reasons, we consider a slight variation of the rule in that we stop at step $\widehat{m}-1$ instead of $\widehat{m}$ if $q_{\widehat{m}}(0) \geq 4\kappa\sqrt{\log(2\gamma^{-1})/n}$, where q_m is the iteration polynomial such that $\alpha_m = q_m(K_n)\Upsilon$. Denote $\widetilde{m}$ the resulting stopping step. We obtain the following result.

Theorem 2.1. *Suppose that Y is bounded* **(Bounded)***, and that the source condition* **SC**(r, ρ) *holds for $r \geq 1/2$. With probability $1 - 2\gamma$, the estimator $f_{\widetilde{m}}$ obtained by the (modified) discrepancy stopping rule* (7) *satisfies*

$$\|f_{\widetilde{m}} - f^*\|_2^2 \leq c(r,\tau)(M+\rho)^2\left(\frac{\log^2\gamma^{-1}}{n}\right)^{\frac{2r}{2r+1}}.$$

We present the proof in Section 4.

2.2 Optimal rates in inner case

We now introduce a stopping rule yielding order-optimal convergence rates as a function of the two parameters r and s in the "inner" case ($r \geq 1/2$, which is equivalent to saying that the target function belongs to $\mathcal{H}$ almost surely). For some constant $\tau' > 3/2$ and $1 > \gamma > 0$, we consider the discrepancy stopping rule with the fixed threshold

$$\Lambda_m \equiv \Lambda = \tau' M\sqrt{\kappa}\left(\frac{4D}{\sqrt{n}}\log\frac{6}{\gamma}\right)^{\frac{2r+1}{2r+s}}. \tag{8}$$

for which we obtain the following:

Theorem 2.2. *Suppose that the noise fulfills the Bernstein assumption* **(Bernstein)***, that the source condition* **SC**(r, ρ) *holds for $r \geq 1/2$, and that* **ED**(s, D) *holds. With probability $1 - 3\gamma$, the estimator $f_{\widehat{m}}$ obtained by the discrepancy stopping rule* (8) *satisfies*

$$\|f_{\widehat{m}} - f^*\|_2^2 \leq c(r,\tau')(M+\rho)^2\left(\frac{16D^2}{n}\log^2\frac{6}{\gamma}\right)^{\frac{2r}{2r+s}}.$$

Due to space limitations, the proof is presented in the supplementary material.

2.3 Optimal rates in outer case, given additional unlabeled data

We now turn to the "outer" case. In this case, we make the additional assumption that *unlabeled* data is available. Assume that we have $\tilde{n}$ i.i.d. observations $X_1, \ldots, X_{\tilde{n}}$, out of which only the first n are labeled. We define a new response vector $\tilde{\Upsilon} = \frac{\tilde{n}}{n}(Y_1, \ldots, Y_n, 0, \ldots, 0) \in \mathbb{R}^{\tilde{n}}$ and run the CG algorithm 1 on $X_1, \ldots, X_{\tilde{n}}$ and $\tilde{\Upsilon}$. We use the same threshold (8) as in the previous section for the stopping rule, except that the factor M is replaced by $\max(M, \rho)$.

Theorem 2.3. *Suppose assumptions* **(Bounded)***,* **SC**(r, ρ) *and* **ED**(s, D)*, with $r + s \geq \frac{1}{2}$. Assume unlabeled data is available with*

$$\frac{\widetilde{n}}{n} \geq \left(\frac{16D^2}{n}\log^2\frac{6}{\gamma}\right)^{-\frac{(1-2r)_+}{2r+s}}.$$

Then with probability $1 - 3\gamma$*, the estimator* $f_{\widehat{m}}$ *obtained by the discrepancy stopping rule defined above satisfies*

$$\|f_{\widehat{m}} - f^*\|_2^2 \leq c(r, \tau')(M + \rho)^2 \left(\frac{16D^2}{n} \log^2 \frac{6}{\gamma} \right)^{\frac{2r}{2r+s}} .$$

A sketch of the proof can be found in the supplementary material.

3 Discussion and comparison to other results

For the inner case – i.e. $f^* \in \mathcal{H}$ almost surely – we provide two different consistent stopping criteria. The first one (Section 2.1) is oblivious to the effective dimension parameter s, and the obtained bound corresponds to the "worst case" with respect to this parameter (that is, $s = 1$). However, an interesting feature of stopping rule (7) is that the rule itself does not depend on the a priori knowledge of the regularity parameter r, while the achieved learning rate does (and with the optimal dependence in r when $s = 1$). Hence, Theorem 2.1 implies that the obtained rule is automatically *adaptive* with respect to the regularity of the target function. This contrasts with the results obtained in [1] for linear regularization schemes of the form (3), (also in the case $s = 1$) for which the choice of the regularization parameter λ leading to optimal learning rates required the knowledge or r beforehand.

When taking into account also the effective dimensionality parameter s, Theorem 2.2 provides the order-optimal convergence rate in the inner case (up to a log factor). A noticeable difference to Theorem 2.1 however, is that the stopping rule is no longer adaptive, that is, it depends on the *a priori* knowledge of parameters r and s. We observe that previously obtained results for linear regularization schemes of the form (2) in [6] and of the form (3) in [5], also rely on the a priori knowledge of r and s to determine the appropriate regularization parameter λ.

The outer case – when the target function does not lie in the reproducing Kernel Hilbert space $\mathcal{H}$ – is more challenging and to some extent less well understood. The fact that additional assumptions are made is not a particular artefact of CG methods, but also appears in the studies of other regularization techniques. Here we follow the semi-supervised approach that is proposed in e.g. [5] (to study linear regularization of the form (3)) and assume that we have sufficient additional unlabeled data in order to ensure learning rates that are optimal as a function of the number of *labeled* data. We remark that other forms of additional requirements can be found in the recent literature in order to reach optimal rates. For regularized M-estimation schemes studied in [20], availability of unlabeled data is not required, but a condition is imposed of the form $\|f\|_\infty \leq C \|f\|_{\mathcal{H}}^p \|f\|_2^{1-p}$ for all $f \in \mathcal{H}$ and some $p \in (0, 1]$. In [13], assumptions on the supremum norm of the eigenfunctions of the kernel integral operator are made (see [20] for an in-depth discussion on this type of assumptions).

Finally, as explained in the introduction, the term 'conjugate gradients' comprises a class of methods that approximate the solution of linear equations on Krylov subspaces. In the context of learning, our approach is most closely linked to Partial Least Squares (PLS) [21] and its kernel extension [18]. While PLS has proven to be successful in a wide range of applications and is considered one of the standard approaches in chemometrics, there are only few studies of its theoretical properties. In [8, 14], consistency properties are provided for linear PLS under the assumption that the target function f^* depends on a finite known number of orthogonal latent components. These findings were recently extended to the nonlinear case and without the assumption of a latent components model [3], but all results come without optimal rates of convergence. For the slightly different CG approach studied by Ong et al [17], bounds on the difference between the empirical risks of the CG approximation and of the target function are derived in [16], but no bounds on the generalization error were derived.

4 Proofs

Convergence rates for regularization methods of the type (2) or (3) have been studied by casting kernel learning methods into the framework of *inverse problems* (see [9]). We use this framework for the present results as well, and recapitulate here some important facts.

We first define the *empirical evaluation operator* T_n as follows:

$$T_n : \qquad g \in \mathcal{H} \mapsto T_n g := (g(X_1), \ldots, g(X_n))^\top \in \mathbb{R}^n$$

and the *empirical integral operator* T_n^* as:

$$T_n^* : u = (u_1, \ldots, u_n) \in \mathbb{R}^n \mapsto T_n^* u := \frac{1}{n} \sum_{i=1}^n u_i k(X_i, \cdot) \in \mathcal{H}.$$

Using the reproducing property of the kernel, it can be readily checked that T_n and T_n^* are adjoint operators, i.e. they satisfy $\langle T_n^* u, g \rangle_{\mathcal{H}} = \langle u, T_n g \rangle$, for all $u \in \mathbb{R}^n, g \in \mathcal{H}$. Furthermore, $K_n = T_n T_n^*$, and therefore $\|\alpha\|_{K_n} = \|f_\alpha\|_{\mathcal{H}}$. Based on these facts, equation (5) can be rewritten as

$$f_m = \arg \min_{f \in \mathcal{K}_m(T_n^* \Upsilon, S_n)} \|T_n^* Y - S_n f\|_{\mathcal{H}}\,, \tag{9}$$

where $S_n = T_n^* T_n$ is a self-adjoint operator of $\mathcal{H}$, called empirical covariance operator. This definition corresponds to that of the "usual" conjugate gradient algorithm formally applied to the so-called normal equation (in $\mathcal{H}$)

$$S_n f_\alpha = T_n^* \Upsilon\,,$$

which is obtained from (1) by left multiplication by T_n^*. The advantage of this reformulation is that it can be interpreted as a "perturbation" of a *population, noiseless* version (of the equation and of the algorithm), wherein Y is replaced by the target function f^* and the empirical operator T_n^*, T_n are respectively replaced by their population analogues, the kernel integral operator

$$T^* : g \in L_2(P_X) \mapsto T^* g := \int k(.,x) g(x) dP_X(x) = \mathbb{E}\left[k(X, \cdot) g(X)\right] \in \mathcal{H}\,,$$

and the change-of-space operator

$$T : \qquad g \in \mathcal{H} \mapsto g \in \mathcal{L}_2(P_X)\,.$$

The latter maps a function to itself but between two Hilbert spaces which differ with respect to their geometry – the inner product of $\mathcal{H}$ being defined by the kernel function k, while the inner product of $\mathcal{L}_2(P_X)$ depends on the data generating distribution (this operator is well defined: since the kernel is bounded, all functions in $\mathcal{H}$ are bounded and therefore square integrable under any distribution P_X).

The following results, taken from [1] (Propositions 21 and 22) quantify more precisely that the empirical covariance operator $S_n = T_n^* T_n$ and the empirical integral operator applied to the data, $T_n^* \Upsilon$, are close to the population covariance operator $S = T^* T$ and to the kernel integral operator applied to the noiseless target function, $T^* f^*$ respectively.

Proposition 4.1. *Assume that $k(x,x) \leq \kappa$ for all $x \in \mathcal{X}$. Then the following holds:*

$$\mathbb{P}\left[\|S_n - S\|_{HS} \leq \frac{4\kappa}{\sqrt{n}} \sqrt{\log \frac{2}{\gamma}}\right] \geq 1 - \gamma\,, \tag{10}$$

where $\|.\|_{HS}$ denotes the Hilbert-Schmidt norm. If the representation $f^* = T f_{\mathcal{H}}^*$ holds, and under assumption **(Bernstein)***, we have the following:*

$$\mathbb{P}\left[\|T_n^* \mathbf{Y} - S f_{\mathcal{H}}^*\| \leq \frac{4M\sqrt{\kappa}}{\sqrt{n}} \log \frac{2}{\gamma}\right] \geq 1 - \gamma\,. \tag{11}$$

We note that $f^* = T f_{\mathcal{H}}^*$ implies that the target function f^* coincides with a function $f_{\mathcal{H}}^*$ belonging to $\mathcal{H}$ (remember that T is just the change-of-space operator). Hence, the second result (11) is valid for the case with $r \geq 1/2$, but it is not true in general for $r < 1/2$.

4.1 Nemirovskii's result on conjugate gradient regularization rates

We recall a sharp result due to Nemirovskii [15] establishing convergence rates for conjugate gradient methods in a deterministic context. We present the result in an abstract context, then show how, combined with the previous section, it leads to a proof of Theorem 2.1. Consider the linear equation

$$Az^* = b\,,$$

where A is a bounded linear operator over a Hilbert space $\mathcal{H}$. Assume that the above equation has a solution and denote z^* its minimal norm solution; assume further that a self-adjoint operator $\bar{A}$, and an element $\bar{b} \in \mathcal{H}$ are known such that

$$\left\|A - \bar{A}\right\| \leq \delta\,; \qquad \left\|b - \bar{b}\right\| \leq \varepsilon\,, \tag{12}$$

(with δ and ε known positive numbers). Consider the CG algorithm based on the noisy operator $\bar{A}$ and data $\bar{b}$, giving the output at step m

$$z_m = \underset{z \in \mathcal{K}_m(\bar{A},\bar{b})}{\operatorname{Arg\,Min}} \left\|\bar{A}z - \bar{b}\right\|^2 . \tag{13}$$

The *discrepancy principle* stopping rule is defined as follows. Consider a fixed constant $\tau > 1$ and define

$$\bar{m} = \min\left\{m \geq 0 : \left\|\bar{A}z_m - \bar{b}\right\| < \tau(\delta \left\|z_m\right\| + \varepsilon)\right\} .$$

We output the solution obtained at step $\max(0, \bar{m} - 1)$. Consider a minor variation of of this rule:

$$\widehat{m} = \begin{cases} \bar{m} & \text{if } q_{\bar{m}}(0) < \eta\delta^{-1} \\ \max(0, \bar{m} - 1) & \text{otherwise,} \end{cases}$$

where $q_{\bar{m}}$ is the degree $m - 1$ polynomial such that $z_{\bar{m}} = q_{\bar{m}}(\bar{A})\bar{b}$, and η is an arbitrary positive constant such that $\eta < 1/\tau$. Nemirovskii established the following theorem:

Theorem 4.2. *Assume that (a)* $\max(\|A\|, \left\|\bar{A}\right\|) \leq L$; *and that (b)* $z^* = A^\mu u^*$ *with* $\|u^*\| \leq R$ *for some* $\mu > 0$. *Then for any* $\theta \in [0, 1]$, *provided that* $\widehat{m} < \infty$ *it holds that*

$$\left\|A^\theta \left(z_{\widehat{m}} - z^*\right)\right\|^2 \leq c(\mu, \tau, \eta) R^{\frac{2(1-\theta)}{1+\mu}} (\varepsilon + \delta R L^\mu)^{2(\theta+\mu)/(1+\mu)} .$$

4.2 Proof of Theorem 2.1

We apply Nemirovskii's result in our setting (assuming $r \geq \frac{1}{2}$): By identifying the approximate operator and data as $\bar{A} = S_n$ and $\bar{b} = T_n^* \mathbf{Y}$, we see that the CG algorithm considered by Nemirovskii (13) is exactly (9), more precisely with the identification $z_m = f_m$.

For the population version, we identify $A = S$, and $z^* = f_{\mathcal{H}}^*$ (remember that provided $r \geq \frac{1}{2}$ in the source condition, then there exists $f_{\mathcal{H}}^* \in \mathcal{H}$ such that $f^* = T f_{\mathcal{H}}^*$).

Condition (a) of Nemirovskii's theorem 4.2 is satisfied with $L = \kappa$ by the boundedness of the kernel. Condition (b) is satisfied with $\mu = r - 1/2 \geq 0$ and $R = \kappa^{-r}\rho$, as implied by the source condition **SC**(r, ρ). Finally, the concentration result 4.1 ensures that the approximation conditions (12) are satisfied with probability $1 - 2\gamma$, more precisely with $\delta = \frac{4\kappa}{\sqrt{n}} \sqrt{\log \frac{2}{\gamma}}$ and $\varepsilon = \frac{4M\sqrt{\kappa}}{\sqrt{n}} \log \frac{2}{\gamma}$. (Here we replaced γ in (10) and (11) by $\gamma/2$, so that the two conditions are satisfied simultaneously, by the union bound). The operator norm is upper bounded by the Hilbert-Schmidt norm, so that the deviation inequality for the operators is actually stronger than what is needed.

We consider the discrepancy principle stopping rule associated to these parameters, the choice $\eta = 1/(2\tau)$, and $\theta = \frac{1}{2}$, thus obtaining the result, since

$$\left\|A^{\frac{1}{2}} \left(z_{\widehat{m}} - z^*\right)\right\|^2 = \left\|S^{\frac{1}{2}} \left(f_{\widehat{m}} - f_{\mathcal{H}}^*\right)\right\|_{\mathcal{H}}^2 = \left\|f_{\widehat{m}} - f_{\mathcal{H}}^*\right\|_2^2 .$$

4.3 Notes on the proof of Theorems 2.2 and 2.3

The above proof shows that an application of Nemirovskii's fundamental result for CG regularization of inverse problems under deterministic noise (on the data and the operator) allows us to obtain our first result. One key ingredient is the concentration property 4.1 which allows to bound deviations in a quasi-deterministic manner.

To prove the sharper results of Theorems 2.2 and 2.3, such a direct approach does not work unfortunately, and a complete rework and extension of the proof is necessary. The proof of Theorem 2.2 is presented in the supplementary material to the paper. In a nutshell, the concentration result 4.1 is too coarse to prove the optimal rates of convergence taking into account the effective dimension

parameter. Instead of that result, we have to consider the deviations from the mean in a "warped" norm, i.e. of the form $\left\|(S+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y}-T^*f^*)\right\|$ for the data, and $\left\|(S+\lambda I)^{-\frac{1}{2}}(S_n-S)\right\|_{HS}$ for the operator (with an appropriate choice of $\lambda > 0$) respectively. Deviations of this form were introduced and used in [5, 6] to obtain sharp rates in the framework of Tikhonov's regularization (2) and of the more general linear regularization schemes of the form (3). Bounds on deviations of this form can be obtained via a Bernstein-type concentration inequality for Hilbert-space valued random variables.

On the one hand, the results concerning linear regularization schemes of the form (3) do not apply to the nonlinear CG regularization. On the other hand, Nemirovskii's result does not apply to deviations controlled in the warped norm. Moreover, the "outer" case introduces additional technical difficulties. Therefore, the proofs for Theorems 2.2 and 2.3, while still following the overall fundamental structure and ideas introduced by Nemirovskii, are significantly different in that context. As mentioned above, we present the complete proof of Theorem 2.2 in the supplementary material and a sketch of the proof of Theorem 2.3.

5 Conclusion

In this work, we derived early stopping rules for kernel Conjugate Gradient regression that provide optimal learning rates to the true target function. Depending on the situation that we study, the rates are adaptive with respect to the regularity of the target function in some cases. The proofs of our results rely most importantly on ideas introduced by Nemirovskii [15] and further developed by Hanke [11] for CG methods in the deterministic case, and moreover on ideas inspired by [5, 6].

Certainly, in practice, as for a large majority of learning algorithms, cross-validation remains the standard approach for model selection. The motivation of this work is however mainly theoretical, and our overall goal is to show that from the learning theoretical point of view, CG regularization stands on equal footing with other well-studied regularization methods such as kernel ridge regression or more general linear regularization methods (which includes between many others L_2 boosting). We also note that theoretically well-grounded model selection rules can generally help cross-validation in practice by providing a well-calibrated parametrization of regularizer functions, or, as is the case here, of thresholds used in the stopping rule.

One crucial property used in the proofs is that the proposed CG regularization schemes can be conveniently cast in the reproducing kernel Hilbert space $\mathcal{H}$ as displayed in e.g (9). This reformulation is not possible for Kernel Partial Least Squares: It is also a CG type method, but uses the standard Euclidean norm instead of the K_n-norm used here. This point is the main technical justification on why we focus on (5) rather than kernel PLS. Obtaining optimal convergence rates also valid for Kernel PLS is an important future direction and should build on the present work.

Another important direction for future efforts is the derivation of stopping rules that do not depend on the confidence parameter γ. Currently, this dependence prevents us to go from convergence in high probability to convergence in expectation, which would be desirable. Perhaps more importantly, it would be of interest to find a stopping rule that is adaptive to both parameters r (target function regularity) and s (effective dimension parameter) without their a priori knowledge. We recall that our first stopping rule is adaptive to r but at the price of being worst-case in s. In the literature on linear regularization methods, the optimal choice of regularization parameter is also non-adaptive, be it when considering optimal rates with respect to r only [1] or to both r and s [5]. An approach to alleviate this problem is to use a hold-out sample for model selection; this was studied theoretically in [7] for linear regularization methods (see also [4] for an account of the properties of hold-out in a general setup). We strongly believe that the hold-out method will yield theoretically founded adaptive model selection for CG as well. However, hold-out is typically regarded as inelegant in that it requires to throw away part of the data for estimation. It would be of more interest to study model selection methods that are based on using the whole data in the estimation phase. The application of Lepskii's method is a possible step towards this direction.

References

[1] F. Bauer, S. Pereverzev, and L. Rosasco. On Regularization Algorithms in Learning Theory. *Journal of Complexity*, 23:52–72, 2007.

[2] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence Rates of General Regularization Methods for Statistical Inverse Problems and Applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636, 2007.

[3] G. Blanchard and N. Krämer. Kernel Partial Least Squares is Universally Consistent. *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop & Conference Proceedings*, 9:57–64, 2010.

[4] G. Blanchard and P. Massart. Discussion of V. Koltchinskii's "Local Rademacher complexities and oracle inequalities in risk minimization". *Annals of Statistics*, 34(6):2664–2671, 2006.

[5] A. Caponnetto. Optimal Rates for Regularization Operators in Learning Theory. Technical Report CBCL Paper 264/ CSAIL-TR 2006-062, Massachusetts Institute of Technology, 2006.

[6] A. Caponnetto and E. De Vito. Optimal Rates for Regularized Least-squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[7] A. Caponnetto and Y. Yao. Cross-validation based Adaptation for Regularization Operators in Learning Theory. *Analysis and Applications*, 8(2):161–183, 2010.

[8] H. Chun and S. Keles. Sparse Partial Least Squares for Simultaneous Dimension Reduction and Variable Selection. *Journal of the Royal Statistical Society B*, 72(1):3–25, 2010.

[9] E. De Vito, L. Rosasco, A. Caponnetto, U. De Giovannini, and F. Odone. Learning from Examples as an Inverse Problem. *Journal of Machine Learning Research*, 6(1):883, 2006.

[10] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[11] M. Hanke. *Conjugate Gradient Type Methods for Linear Ill-posed Problems*. Pitman Research Notes in Mathematics Series, 327, 1995.

[12] L. Lo Gerfo, L. Rosasco, E. Odone, F.and De Vito, and A. Verri. Spectral Algorithms for Supervised Learning. *Neural Computation*, 20:1873–1897, 2008.

[13] S. Mendelson and J. Neeman. Regularization in Kernel Learning. *The Annals of Statistics*, 38(1):526–565, 2010.

[14] P. Naik and C.L. Tsai. Partial Least Squares Estimator for Single-index Models. *Journal of the Royal Statistical Society B*, 62(4):763–771, 2000.

[15] A. S. Nemirovskii. The Regularizing Properties of the Adjoint Gradient Method in Ill-posed Problems. *USSR Computational Mathematics and Mathematical Physics*, 26(2):7–16, 1986.

[16] C. S. Ong. Kernels: Regularization and Optimization. *Doctoral dissertation, Australian National University*, 2005.

[17] C. S. Ong, X. Mary, S. Canu, and A. J. Smola. Learning with Non-positive Kernels. In *Proceedings of the 21st International Conference on Machine Learning*, pages 639 – 646, 2004.

[18] R. Rosipal and L.J. Trejo. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Spaces. *Journal of Machine Learning Research*, 2:97–123, 2001.

[19] R. Rosipal, L.J. Trejo, and B. Matthews. Kernel PLS-SVC for Linear and Nonlinear Classification. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 640–647, Washington, DC, 2003.

[20] I. Steinwart, D. Hush, and C. Scovel. Optimal Rates for Regularized Least Squares Regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.

[21] S. Wold, H. Ruhe, H. Wold, and W.J. Dunn III. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal of Scientific and Statistical Computations*, 5:735–743, 1984.

[22] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.