Sharp analysis of low-rank kernel matrix approximations

Francis Bach
INRIA - Sierra project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France
francis.bach@ens.fr

August 9, 2012

Abstract

We consider supervised learning problems within the positive-definite kernel framework, such as kernel ridge regression, kernel logistic regression or the support vector machine. With kernels leading to infinite-dimensional feature spaces, a common practical limiting difficulty is the necessity of computing the kernel matrix, which most frequently leads to algorithms with running time at least quadratic in the number of observations n, i.e., $O(n^2)$. Low-rank approximations of the kernel matrix are often considered as they allow the reduction of running time complexities to $O(p^2 n)$, where p is the rank of the approximation. The practicality of such methods thus depends on the required rank p. In this paper, we show that for approximations based on a random subset of columns of the original kernel matrix, the rank p may be chosen to be linear in the *degrees of freedom* associated with the problem, a quantity which is classically used in the statistical analysis of such methods, and is often seen as the implicit number of parameters of non-parametric estimators. This result enables simple algorithms that have sub-quadratic running time complexity, but provably exhibit the same *predictive performance* than existing algorithms.

1 Introduction

Kernel methods, such as the support vector machine or kernel ridge regression, are now widely used in many areas of science and engineering, such as computer vision or bioinformatics (see, e.g., [1, 2, 3, 4]). Their main attractive features are that (1) they allow non-linear predictions through the same algorithms than for linear predictions, owing to the kernel trick; (2) they allow the separation of the representation problem (designing good kernels for non-vectorial data) and the algorithmic/theoretical problems (given a kernel, how to design, run efficiently and analyze estimation algorithms). Moreover, (3) their applicability goes beyond supervised learning problems, through the kernelization of classical unsupervised learning techniques such as principal component analysis or K-means. Finally, (4) probabilistic Bayesian interpretations through Gaussian processes allow their simple use within larger probabilistic models. For more details, see, e.g., [5, 1, 2].

However, kernel methods typically suffer from at least quadratic running-time complexity in the number of observations n, as this is the complexity of computing the kernel matrix. In large-scale settings where n may be large, this is usually not acceptable. In these situations where plain kernel methods cannot be run, practitioners would commonly (a) turn to methods such as boosting, decision trees or random forests, which have both good running time complexity and predictive performance. However, these methods are typically run on data coming as vectors and usually put a strong emphasis on a sequence of decisions based on single variables. Another common solution is (b) to stop using infinite-dimensional kernels and restrict the kernels to be essentially linear kernels (i.e., by choosing an explicit representation of the data whose size is independent of the number of observations) where the non-parametric kernel machinery (of adapting the complexity of the underlying predictor to the size of the dataset) is lost, and the methods may then *underfit*.

In this paper, we consider the traditional kernel set-up for supervised learning, where the input data are only known through (portions of) the kernel matrix. The main question we try to tackle is the following: Is it possible to run supervised learning methods with positive-definite kernels in time which is subquadratic in the number of observations without losing prediction performance? Of course, if adaptation is desired, linear complexity seems impossible, and therefore we should expect (hopefully slightly) super-linear algorithms. Statistically, a quantity that characterizes the non-parametric nature of kernel method is the *degrees of freedom*, which play the role of an implicit number of parameters and which we define in Section 4.1. Does it play a role in the computational properties of kernel methods?

An important feature of kernel matrices is that they are positive-semidefinite, and thus they may well be approximated from a random subset of p of their columns, in running-time complexity $O(p^2 n)$ and with a computable bound on the error (see details in Section 3). This appears through different formulations within numerical linear algebra or machine learning, e.g., Nyström method [6], sparse greedy approximations [7], incomplete Cholesky decomposition [8], Gram-Schmidt orthonormalization [2] or CUR matrix decompositions [9]. It has been studied a lot [9, 10, 11], in contexts where the goal is kernel matrix approximation or approximate eigenvalue decomposition. Such bounds have also been subsequently used to characterize the approximation of predictions made from these low-rank decompositions [12, 13], but this two-stage analyses do not lead to guarantees that reflect the good observed practical behavior. In this paper, our analysis aims at answering explicitly the simple question: how big should p be to incur no loss of predictive performance compared to the full kernel matrix? The key insight of this paper is not to try to approximate the kernel matrix well, but to predict well from the approximation. This requires a sharper analysis of the approximation properties of the column sampling approach.

We make the following contributions:

- In the least-squares regression setting, we show in Section 4.2 that the rank p can be chosen to be linear in the *degrees of freedom* associated with the problem.
- We present in Section 4.4 simple algorithms that have sub-quadratic running time complexity, and, for the square loss, provably exhibit the same *predictive performance* as classical algorithms than run in quadratic time (or more).
- We provide in Section 4.3 explicit examples of optimal values of the regularization parameters, as a function of the decay of the eigenvalues of the kernel matrix, shedding some light in the joint computational/statistical trade-offs for choosing a good kernel. In particular, we show that with kernels with fast spectrum decays (such as the exponential or Gaussian kernels), com-

putational limitations may prevent exploring the relevant portions of the regularization paths, leading to underfitting.

2 Supervised learning with positive-definite kernels

In this section, we present the problem we try to solve, as well as several areas of the machine learning and statistics literatures our method relates to.

2.1 Equivalent formulations

Let (x_i, y_i), $i = 1, \ldots, n$, be n pairs of points in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space, and $\mathcal{Y}$ is the set of outputs/labels. In this paper, we consider the problem of minimizing

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|^2, \tag{1}$$

where $\mathcal{F}$ is a reproducing kernel Hilbert space with feature map $\phi : \mathcal{X} \to \mathcal{F}$, and positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. While this problem is formulated as an optimization problem in a Hilbert space, it may be formulated as the optimization over $\mathbb{R}^n$ in two different ways.

First, using the representer theorem (see, e.g., [1, 2]), the unique solution f may be found as $f = \sum_{i=1}^{n} \alpha_i \phi(x_i)$. Thus, by replacing the expression of f in Eq. (1), α is a solution of the following optimization problem:

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha, \tag{2}$$

where $K \in \mathbb{R}^{n \times n}$ is the *kernel matrix*, defined as $K_{ij} = k(x_i, x_j)$.

Second, for convex losses only, an equivalent dual problem is classically obtained as (see proof in the appendix):

$$\max_{\alpha \in \mathbb{R}^n} -g(-\lambda\alpha) - \frac{\lambda}{2} \alpha^\top K \alpha, \tag{3}$$

where $g(z) = \max_{u \in \mathbb{R}^n} -\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, u_i) + u_i z_i$ is the Fenchel-conjugate of the empirical risk (for the hinge loss, Eq. (3) is exactly the classical dual formulation of the SVM). Again, one may express the primal solution as $f = \sum_{i=1}^{n} \alpha_i \phi(x_i)$. In many situations (such as with the square loss or logistic loss), then the solution of Eq. (3) is unique, and it is also a solution of Eq. (2) (note however that the converse is not true).

2.2 Related work

Efficient optimization algorithms for kernel methods. In order to solve Eq. (1), algorithms typically consider a primal or a dual approach. Solving Eq. (2) (i.e., primal formulation after application of the representer theorem) is typically inefficient because the problem is ill-conditioned[1] and thus second-order algorithms are typically used [14]. Alternatively, K is represented explicitly as $K = \Phi\Phi^\top$ and a change of variable $w = \Phi^\top\alpha$ is considered (note that when the kernel k is linear, Φ is simply the design matrix, and we are solving directly a linear supervised learning problem). Then, the classical battery of convex optimization algorithms may be used, such as gradient descent, stochastic gradient descent [15] or cutting-planes [16]. However, in a kernel setting where a small matrix Φ (i.e., with few columns) is not known a priori, then they all exhibit at least quadratic complexity in n, as the full kernel matrix is used.

The dual problem in Eq. (3) is usually better-behaved (it has a better condition number) [14], and algorithms such as coordinate descent and its variants such as sequential minimal optimization may be used [17]. Again, the full kernel matrix is needed.

Some algorithms do not need to compute the full kernel matrix, such as the "forgetron" [18] or the "projectron" [19], which operate online on a fixed budget and come with theoretical approximation guarantees; however, these do not characterize the required rank which is needed to achieve the same accuracy than the problem with a full kernel matrix. In fact, one of the main motivations for this work is to derive precise bounds for reduced-set stochastic gradient algorithms for supervised kernel problems.

Analysis of column sampling approximation. Given a positive semi-definite matrix K of size n, many methods exist for approximating it with a low-rank (typically also positive semidefinite) matrix L. While the optimal approximation is obtained from the eigenvalue decomposition, it is not computationally efficient as it has complexity at least quadratic in n (as it requires the knowledge of K). In order to achieve linear complexity in n, approximations from subsets of columns are considered and appear under many names: Nyström method [6], sparse greedy approximations [7], incomplete Cholesky decomposition [8], Gram-Schmidt orthonormalization [2] or CUR matrix decompositions [9]. Note that reduced-set methods (see, e.g., [20]) typically consider using a subset of columns after the predictor has been estimated. These low-rank methods are described in Section 3 and have running time complexity $O(p^2n)$ for an approximation of rank p. Note that they may also be used in a Bayesian setting with Gaussian processes (see, e.g., [21]).

Column sampling has been analyzed a lot [9, 12, 11]; however, typically the analysis provides a bound on the error $\|K - L\|$ for an appropriate norm (typically operator, Frobenius or trace norm), but this is too pessimistic and does not really match with good practical performance (see more details in Section 4.2). Some works do consider prediction guarantees [12, 13], but as shown in Section 4.2, these are not sufficient to reach sharp results depending on the degrees of freedom. Moreover, many analyses consider situations where the matrix K is close to low-rank, which is not the case with kernel matrices. In this paper, the control of $K - L$ is more precise and adapted to the use of K within a supervised learning method.

Dimension reduction for linear predictions. The method presented in this paper, which considers random columns from the original kernel matrix, is also related to random projection techniques used for linear prediction problems (see, e.g., [22, 23]). These techniques require however the knowledge of a matrix square root Φ (such that $K = \Phi\Phi^\top$), which leads to complexity greater than

[1]The objective function in Eq. (2) is a function of $K^{1/2}\alpha$, with a kernel matrix K which is often ill-conditioned, usually leading to ill-conditioning of the original problem [14].

quadratic, and then consider replacing it by ΦA where A is a random matrix with fewer columns than rows.

Theoretical analysis of predictive performance of kernel methods. In order to assess the required precision in approximating the kernel matrix, it is key to understand the typical predictive performance of kernel methods. For the square loss, this is classically obtained from a bias-variance decomposition of this performance (see Section 4). A key quantity is the *degrees of freedom*, which play the role of an implicit number of parameters and is applicable to many non-parametric estimation methods which consists in "smoothing" the response vector by a linear operator (see, e.g., [24, 25, 26, 27]). See precise definitions in Section 4.1.

3 Approximation from subset of columns

Approximation from columns. Given a random subset I of $V = \{1, \dots, n\}$ of cardinality p, we simply consider the approximation of the kernel matrix K from the knowledge of $K(V, I)$ (the columns of K indexed by I), by the matrix

$$L = K(V, I)K(I, I)^\dagger K(I, V) = k(x_V, x_I)k(x_I, x_I)^\dagger k(x_I, x_V), \tag{4}$$

where $M^\dagger$ denotes the pseudo-inverse of M, and $k(x_A, x_B)$ denotes the $|A| \times |B|$ matrix composed of elements $k(x_i, x_j)$ for $(i, j) \in A \times B$. This corresponds to creating an explicit feature map of dimension p, i.e., $\widetilde{\phi}(x) = k(x_I, x_I)^{-1/2} k(x_I, x) \in \mathbb{R}^p$, and, this allows the application to test data points (note that using such techniques also allows better *testing* running time peformance).

Given the true feature map $\phi(x) \in \mathcal{F}$, we have $\widetilde{\phi}(x) = k(x_I, x_I)^{-1/2} \phi(x_I)^\top \phi(x) \in \mathbb{R}^p$, and thus we simply perform a linear dimension reduction. Given the fact that we consider random subsets I of size p, this is similar to a random projection, but here the randomness is associated to the specifics of the kernel problem.

Such a feature map may be efficiently obtained in running time $O(p^2 n)$ using incomplete Cholesky decomposition (often interpreted as partial Gram-Schmidt orthonormalization [2]), with the possibility of having a bound on the trace norm of the approximation error (see, e.g., [2]).

Pivoting vs. random sampling. While selecting a random subset is computationally efficient, it may not lead to the best performance. For the task of approximating the kernel matrix, algorithms such as the incomplete Cholesky decomposition *with pivoting*, provide an approximate greedy algorithm with the same complexity than random subsampling [7, 8].

In Section 5, we provide comparisons between the two approaches, showing the potential advantage of the greedy method over random subsampling. However, the analysis of such algorithms is harder, and, to the best of our knowledge, still remains an open problem.

4 Fixed design analysis for least-square regression (ridge regression)

To simplify the analysis, we assume that the n data points $x_1, \dots, x_n$ are deterministic and that $\mathcal{Y} = \mathbb{R}$. In this setting, the classical *generalization error* (prediction error on unseen data points) is replaced by the *in-sample prediction error* (prediction error on observed data points). This fixed design assumption could be relaxed by using tools from [27] (results for random design settings are typically similar to the fixed design settings).

We assume that the loss ℓ is the square loss, i.e., $\ell(y_i, f(x_i)) = \frac{1}{2}(y_i - f(x_i))^2$. By using the representer theorem (see Section 2.1), we classically obtain:

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) \text{ with } \alpha = (K + n\lambda I)^{-1} y.$$

This leads to a prediction vector $\hat{z} = K(K + n\lambda I)^{-1} y \in \mathbb{R}^n$, which is a linear function of the output observations y, and is often referred to as a smoothed estimate of z.

4.1 Analysis of the in-sample prediction error

We denote by $z_i = \mathbb{E} y_i \in \mathbb{R}$ the expectation of y_i, and we denote by $\varepsilon_i = y_i - z_i = y_i - \mathbb{E} y_i \in \mathbb{R}$ the noise variables; they have zero mean and are *only* assumed to have finite covariance matrix C (note that the noise may neither be independent nor identically distributed).

Bias/variance decomposition of the generalization error. Following classical results from the statistics literature (see, e.g., [24, 25, 26]), we obtain the following expected prediction error:

$$\begin{aligned}
\tfrac{1}{n}\mathbb{E}_\varepsilon \|\hat{z} - z\|^2 &= \tfrac{1}{n}\|\mathbb{E}_\varepsilon \hat{z} - z\|^2 + \tfrac{1}{n} \operatorname{tr} \operatorname{var}_\varepsilon(\hat{z}) \\
&= \tfrac{1}{n}\|(I - K(K + n\lambda I)^{-1}) z\|^2 + \tfrac{1}{n} \operatorname{tr} C K^2 (K + n\lambda I)^{-2} \\
&= n\lambda^2 z^\top (K + n\lambda I)^{-2} z + \tfrac{1}{n} \operatorname{tr} C K^2 (K + n\lambda I)^{-2},
\end{aligned}$$

which may be classically decomposed in two terms:

$$\begin{aligned}
\operatorname{bias}(K) &= n\lambda^2 z^\top (K + n\lambda I)^{-2} z \\
\operatorname{variance}(K) &= \tfrac{1}{n} \operatorname{tr} C K^2 (K + n\lambda I)^{-2}.
\end{aligned}$$

Note that the bias term is a matrix-decreasing function of K/λ (and thus an increasing function of λ), while the variance term is a matrix-increasing function of K/λ and the noise covariance matrix C.

Degrees of freedom. Note that an assumption which is usually made is $C = \sigma^2 I$, and the variance term then takes the form $\sigma^2 \operatorname{tr} K^2 (K + n\lambda I)^{-2}$ and $\operatorname{tr} K^2 (K + n\lambda I)^{-2}$ is referred to as the *degrees of freedom* [24, 25, 26, 27] (note that an alternative definition is often used, i.e., $\operatorname{tr} K(K + n\lambda I)^{-1}$, and that as shown in the appendix, they behave similarly). In ordinary least-squares estimation from d variables, the variance term is equal to $\sigma^2 d/n$, and thus the degrees of freedom play the role of an implicit number of parameters. In this paper, we show that a proxy to this statistical quantity also plays a role in optimization: the number of columns needed to approximate the kernel matrix precisely enough to incur no loss of performance is linear in the degrees of freedom.

More precisely, we define the *maximal marginal degrees of freedom* d as

$$d = n \big\| \operatorname{diag}\big(K(K + n\lambda I)^{-1}\big) \big\|_\infty. \tag{5}$$

We have $\operatorname{tr} K^2(K+n\lambda I)^{-2} \leqslant \operatorname{tr} K(K+n\lambda I)^{-1} = \big\| \operatorname{diag}\big(K(K+n\lambda I)^{-1}\big)\big\|_1 \leqslant d$, and thus d provides an upper-bound on the regular degrees of freedom. It may be significantly larger in situations where there may be outliers and the vector $\operatorname{diag}\big(K(K+n\lambda I)^{-1}\big)$ is far from uniform.

4.2 Predictive performance of column sampling

We consider sampling p columns (without replacement) from the original n columns. We consider the column sampling approximation defined in Eq. (4) and provide sufficient conditions (a lower-bound on p) to obtain the same predictive performance than with the full kernel matrix.

Theorem 1 (Generalization performance of column sampling) *Assume $z \in \mathbb{R}^n$ and $K \in \mathbb{R}^{n\times n}$ are respectively a deterministic vector and a symmetric positive semi-definite matrix, and $\lambda > 0$. Let $d = n\big\| \operatorname{diag}\big(K(K+n\lambda I)^{-1}\big)\big\|_\infty$ and $R^2 = \|\operatorname{diag}(K)\|_\infty$. Assume $\varepsilon \in \mathbb{R}^n$ is a random vector with finite variance and zero mean, and define the smoothed estimate $\hat{z}_K = (K+n\lambda I)^{-1}K(z+\varepsilon)$. Assume that I is a uniform random subset of p indices in $\{1,\dots,n\}$ and consider $L = K(V,I)K(I,I)^\dagger K(I,V)$, with the approximate smoothed estimate $\hat{z}_L = (L+n\lambda I)^{-1}L(z+\varepsilon)$. Let $\delta \in (0,1)$. If*

$$p \geqslant \big(\frac{32d}{\delta} + 2\big) \log \frac{nR^2}{\delta\lambda}, \tag{6}$$

then

$$\frac{1}{n}\mathbb{E}_I \mathbb{E}_\varepsilon \|\hat{z}_L - z\|^2 \leqslant \frac{1}{n}(1+4\delta)\mathbb{E}_\varepsilon\|\hat{z}_K - z\|^2. \tag{7}$$

We can make the following observations:

- The bound in Eq. (7) provides a relative approximation guarantee: the predictions $\hat{z}_L$ are shown to perform as well as $\hat{z}_K$ (no kernel matrix approximation). Small values of δ impose no loss of performance, while $\delta = 1/4$ impose that the prediction errors have a similar behavior (up to a factor of 2).
- The lower bound for the rank p in Eq. (6) shows that the maximal marginal degrees of freedom provides a quantity which, up to logarithmic terms, is sufficient to scale with, in order to incur no loss of prediction performance. Note that the previous result also allows the derivation of an approximation guarantee δ given a rank p, by inverting Eq. (6).
- The bound in Eq. (7) provides a result in expectation, both with respect to the data (i.e., $\mathbb{E}_\varepsilon$) and the sampling of columns (i.e., $\mathbb{E}_I$). While results in high-probability with respect to I are readily obtained (in fact, the proof is based on such results), doing so with respect to ε would require additional assumptions, which are standard in the analysis in ridge regression (see, e.g., [27, 28]), but that would make the results significantly more complicated.
- Theorem 1 shows that in the specific instance that we are faced with, we do not lose any average predictive performance. This is different than achieving a good approximation of the kernel matrix [9]. Previous work [12, 13] considers explicitly the use of kernel matrix approximation bounds within classifiers or regressors, but obtain bounds that involve multiplicative terms of the form $1/\lambda^2$, which, as we show in Section 4.3, would grow as n grows. Our proof technique, that focuses *directly* on prediction performance, avoids this, and our dependence is only logarithmic

in λ (see details in the proof in the appendix). Finally, as opposed to [13], the bound is not on the worst-case predictive performance (obtained from optimizing over λ, and with worst-case analysis over K), but for given λ and K.

– Theorem 1 provides a sufficient lower-bound for the required rank p. Deriving precise necessary lower-bounds is outside the scope of this paper. However, given that with a reduced space of p dimensions, we can achieve a prediction error of $O(p/n)$ from ordinary least-squares, we should expect p to be larger than the known minimax rates of estimation for the problem at hand (see, e.g., [29]). In Section 4.3, we show that in some situations, it turns out that d is of the order of the minimax rate; therefore, we could expect that in certain settings, d is also a necessary lower-bound on p (up to constants and logarithms).

– In the existing analysis of sampling techniques for kernel methods, another source of inefficiency which makes our result sharper is the proof technique for bounding $\|K - L\|$. Indeed, most analyses use a linear algebra lemma from [9, 10], that relies on the $(p+1)$-th eigenvalue to be small; hence it is adapted to matrices with sharp eigenvalue decrease, which is not the case for kernel matrices (see an illustrative example in the appendix, where the kernel approximation error decays much slower than the prediction error as the rank p increases). We provide a new proof technique based on regularizing the column sampling approximation and optimizing the extra regularization parameter using a monotonicity argument.

– In our experiments, we have noticed that the low-rank approximation may have an additional regularizing effect leading to a *better* prediction performance than with the full kernel matrix.

4.3 Optimal choice of the regularization parameter

For simplicity, in this section, we assume that the noise variables ε are i.i.d. (i.e., $C = \sigma^2 I$). Our goal is to study simplified situations, where we can derive explicit formulas for the bias, the variance, and the optimal regularization parameter. Throughout this section, we will consider specific decays of certain sequences, which we characterize with the notation $u_n = \Theta(v_n)$, which means that there exist strictly positive constants A and B such that $Au_n \leqslant v_n \leqslant Bv_n$ for all n.

We assume that the kernel matrix K has eigenvalues of the form $\Theta(n\mu_i)$, $i = 1, \dots, n$, for some summable sequence (μ_i)—so that $\operatorname{tr} K = \Theta(n)$, and that the coordinates of z on the eigenbasis of K have the asymptotic behavior $\Theta(\sqrt{n\nu_i})$ for a summable sequence (ν_i)—so that $\frac{1}{n} z^\top z = \Theta(1)$. In Table 1, we provide asymptotic equivalents of all quantities for several pairs of sequences (μ_i) and (ν_i) (see proofs in the appendix), with polynomial or exponential decays.

Note that for decays of ν_i which are polynomial, i.e., $\nu_i = O(i^{-2\delta})$, then the best possible prediction performance is known to be $O(n^{1/2\delta - 1})$ [29] and is achieved if the RKHS is large enough (lines 2 and 4 in Table 1). For exponential decay, the best performance is $O(\log n/n)$.

Given a specific decay (ν_i) for expected outputs $z = \mathbb{E} y$, then depending on the decay (μ_i) of the eigenvalues of the kernel matrix, the final prediction performance and the optimal regularization parameter may be different. Usually, the smaller the RKHS, the faster the decay of eigenvalues of the kernel matrix K (this is true for translation-invariant kernels [1], and the kernels considered in Section 5). Thus there are two regimes:

(μ_i)	(ν_i)	var.	bias	optimal λ	pred. perf.	deg. freed. d	condition
$i^{-2\beta}$	$i^{-2\delta}$	$n^{-1}\lambda^{-1/2\beta}$	λ^2	$n^{-1/(2+1/2\beta)}$	$n^{1/(4\beta+1)-1}$	$n^{1/(4\beta+1)}$	if $2\delta > 4\beta+1$
$i^{-2\beta}$	$i^{-2\delta}$	$n^{-1}\lambda^{-1/2\beta}$	$\lambda^{(2\delta-1)/2\beta}$	$n^{-\beta/\delta}$	$n^{1/(2\delta)-1}$	$n^{1/(2\delta)}$	if $2\delta < 4\beta+1$
$i^{-2\beta}$	$e^{-\kappa i}$	$n^{-1}\lambda^{-1/2\beta}$	λ^2	$n^{-1/(2+1/2\beta)}$	$n^{1/(4\beta+1)-1}$	$n^{1/(4\beta+1)}$	
$e^{-\rho i}$	$i^{-2\delta}$	$n^{-1}\log\frac{1}{\lambda}$	$(\log\frac{1}{\lambda})^{1-2\delta}$	$\exp(-n^{1/(2\delta)})$	$n^{1/(2\delta)-1}$	$n^{1/(2\delta)}$	
$e^{-\rho i}$	$e^{-\kappa i}$	$n^{-1}\log\frac{1}{\lambda}$	λ^2	$n^{-1/2}$	$\log n/n$	$\log n$	if $\kappa > 2\rho$
$e^{-\rho i}$	$e^{-\kappa i}$	$n^{-1}\log\frac{1}{\lambda}$	$\lambda^{\kappa/\rho}$	$n^{-\rho/\kappa}$	$\log n/n$	$\log n$	if $\kappa < 2\rho$

Table 1: Variance, bias, optimal regularization parameter and corresponding prediction performance, for several decays of eigenvalues and signal coefficients (we always assume $\delta > 1/2$, $\beta > 1/2$, $\rho > 0$, $\kappa > 0$, to make the series summable). All entries are functions of i, n or λ and are only asymptotic bounded below and above, i.e., corresponding to the asymptotic notation $\Theta(\cdot)$.

– **The RKHS is too large**, for lines 1 and 3 in Table 1: the eigenvalues of K, which depend linearly on μ_i, do not decay fast enough. In other words, the functions in the RKHS are not smooth enough. In this situation, the prediction performance is suboptimal (do not attain the best possible rate).

– **The RKHS is too small**, for lines 2, 4, and 6 in Table 1: the eigenvalues of K decay fast enough to get an optimal prediction performance. In other words, the functions in the RKHS are potentially smoother than what is necessary. In this situation however, the required value of λ may be very small (much smaller than $O(n^{-1})$), leading to potentially harder optimization problems (since the condition number that depends on $1/\lambda$ may be very large).

There is thus a computational/statistical trade-off: if the RKHS is chosen too large, then the prediction performance is suboptimal; if the RKHS is chosen too small, the prediction performance could be optimal, but the optimization problems are harder, and sometimes cannot be solved with the classical precision of numerical techniques (see examples of such behavior in Section 5).

4.4 Optimization algorithms with column sampling

Given a rank p and a regularization parameter λ, we consider the following algorithm to solve Eq. (1) for twice differentiable convex losses:

1. Select at random p columns of K (without replacement).
2. Compute $\Phi \in \mathbb{R}^{n\times p}$ such that $\Phi\Phi^\top = K(V,I)K(I,I)^\dagger K(I,V)$ using incomplete Cholesky decomposition (see details in [2]).
3. Minimize $\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \ell(y_i,(\Phi w)_i) + \frac{\lambda}{2}\|w\|^2$ using Newton's method (i.e., a single linear system for the square loss).

The complexity of step 2 is already $O(p^2 n)$, therefore using faster techniques for step 3 (e.g., accelerated gradient descent) does not change the overall complexity, which is thus $O(p^2 n)$. Moreover, since we use a second-order method for step 3, we are robust to ill-conditioning and in particular to small values of λ (though not below machine precision as seen in Section 5). This is not the case for algorithms that relies on the strong convexity of the objective function, whose convergence is much slower when λ is small (as seen in Section 4.3, when n grows, the optimal value of λ can decay very rapidly, making these traditional methods non robust).

According to Theorem 1, at least for the square loss, the dimension p may be chosen to be linear in the degrees of freedom d, which, as illustrated in Table 1, is typically smaller than $n^{1/2}$ (d is of the order of the prediction performance multiplied by n). Therefore, if p is properly chosen, the complexity is subquadratic. Given λ, d (and thus p) can be estimated from a low-rank approximation of K. However, our current analysis assumes that λ is given. Selecting the rank p *and* the regularization parameter λ in a data-driven way would make the prediction method more robust, but this would require extra assumptions (see, e.g., [28] and references therein).

5 Simulations

Synthetic examples. In order to study various behaviors of the regularization parameters λ and the degrees of freedom d, we consider periodic smoothing splines on $[0, 1]$ and points $x_1, \dots, x_n$ uniformly spread over $[0, 1]$, either deterministically or randomly. In order to generate problems with given sequences (μ_i) and (ν_i), it suffices to choose $k(x, y) = \sum_{i=1}^{\infty} 2\mu_i \cos 2i\pi(x - y)$, and a function $f(x) = \sum_{i=1}^{\infty} 2\nu_i^{1/2} \cos 2i\pi x$. For $\mu_i = i^{-2\beta}$, we have $k(x, y) = \frac{1}{(2\beta)!} B_{2\beta}(x-y-\lfloor x-y \rfloor)$, where $B_{2\beta}$ is the (2β)-th Bernoulli polynomial (see details in the appendix).

Optimal values of λ. In a first experiment, we illustrate the results from Section 4.3, and compute in Figure 1 the best value of the regularization parameter (left) and the obtained predictive performance (middle), for a problem with $\nu_i = i^{-2\delta}$ for $\delta = 8$, and for which we considered several kernels, for which $\mu_i = i^{-2\beta}$, for $\beta = 1$, $\beta = 4$ and $\beta = 8$. We can make the following observations:

– For $\beta = 1$, the rate of convergence of $n^{1/(4\beta+1)-1}$ is achieved (line 1 in Table 1), with a certain asymptotic decay of the regularization parameter, and it is slower than $n^{1/(2\delta)-1}$.

– For $\beta = 4$, the optimal rate of $n^{1/(2\delta)-1}$ is achieved (line 2 in Table 1), as expected.

– For $\beta = 8$, the rate of convergence should be $n^{1/(2\delta)-1}$ (line 2 in Table 1), however, as seen in the left plot, the regularization parameter saturates as n grows at the machine precision, leading, because of numerical errors, to worse prediction performance. The problem is so ill-conditioned that the matrix inversion cannot be algorithmically robust enough.

Performance of low-rank approximations. In this series of experiments, we compute the rank p which is necessary to achieve a predictive performance at most 1% worse than with $p = n$, and compute[2] the ratio with the marginal degrees of freedom $d = n\big\| \operatorname{diag}\big(K(K + n\lambda I)^{-1}\big)\big\|_\infty$ and the traditional degrees of freedom $d_{ave} = \operatorname{tr} K^2(K + n\lambda I)^{-2}$. In the right plot of Figure 1, we consider data randomly distributed in $[0, 1]$ with the same kernels and functions than above, while

[2]Note that in practice, computing the degrees of freedom exactly requires to know the full matrix. However, it can also be approximated efficiently using a low-rank approximation based on column sampling.

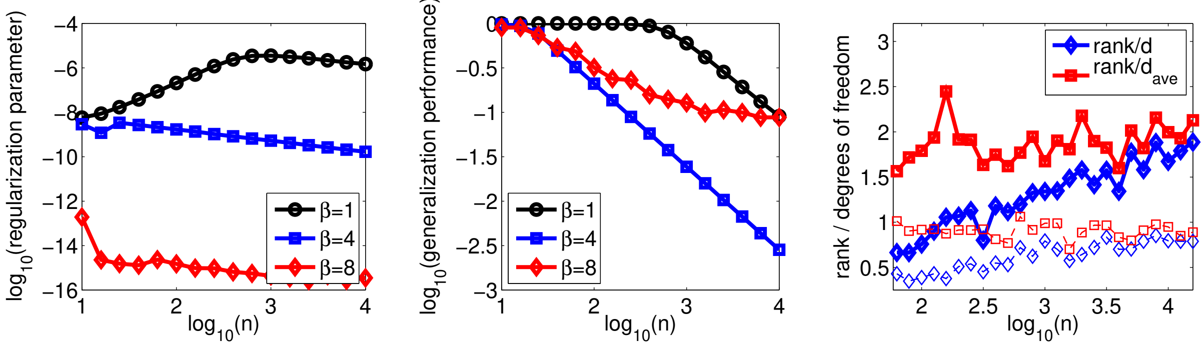

Figure 1: Left and middle: Effect of size of RKHS in predictive performance. Right: Ratio of the sufficient rank to obtain 1% worse predictive performance, over the degrees of freedom (plain: random column sampling, dashed: incomplete Cholesky decomposition with column pivoting).

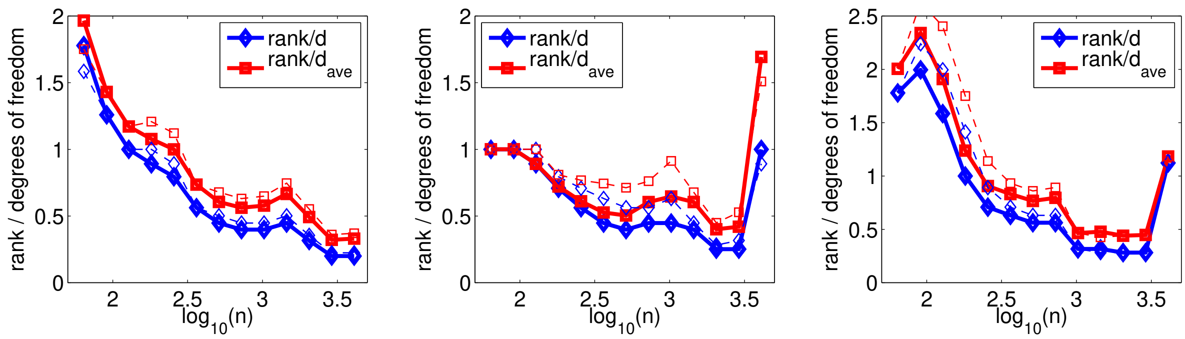

Figure 2: Ratio of the sufficient rank to obtain 1% worse predictive performance, over the degrees of freedom (plain: random column sampling, dashed: incomplete Cholesky decomposition with column pivoting). From left to right: *pumadyn* datasets *32fh*, *32nh*, *32nm*.

in Figure 2, we considered three of the *pumadyn* datasets from the UCI machine learning repository (here we compute the classical generalization performance on unseen data points). We can make the following observations:

– On all datasets, the ratios stay relatively close to one, illustrating the results from Theorem 1.

– Using pivoting to select the columns does not change significantly the results, but may sometimes reduce the number of required columns by a constant factor.

6 Conclusion

In this paper, we have provided an analysis of column sampling for kernel least-squares regression that shows that the rank may be chosen proportional to the degrees of freedom of the problem, showing that the statistical quantity characterizing prediction performance also plays a computational role. The current analysis could be extended in various ways: First, other column sampling schemes beyond uniform, such as presented in [10, 11], could be considered with potentially better behavior; the analysis may also be extended to other losses than the square loss, such as the logistic loss, using self-concordant analysis [30]. Finally, in this paper, we have considered a batch setting and extending these results to online settings is of significant practical and theoretical interest. In particular, the difficulty would be to study regularization parameters that adapt to the number of observations, which lead to vanishing strong convexity constants and does not allow rates of order $O(n^{-1})$.

Acknowledgments

This work was supported by the European Research Council (SIERRA Project).

A Duality for kernel supervised learning

We consider the following problem, where $\mathcal{F}$ is an RKHS with feature map $\phi : \mathcal{X} \to \mathcal{F}$:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \frac{\lambda}{2} \|f\|^2,$$

which may be rewritten with the feature map ϕ as:

$$\min_{f \in \mathcal{F}, u \in \mathbb{R}^n,} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, u_i) + \frac{\lambda}{2} \|f\|^2 \text{ such that } u_i = \langle f, \phi(x_i) \rangle.$$

We may then introduce dual parameters (Lagrange multipliers) $\alpha \in \mathbb{R}^n$ and the Lagrangian

$$\mathcal{L}(f, u, \alpha) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, u_i) + \frac{\lambda}{2} \|f\|^2 + \lambda \sum_{i=1}^{n} \alpha_i (u_i - \langle f, \phi(x_i) \rangle).$$

Minimizing with respect to (f, u), we get $f = \sum_{i=1}^{n} \alpha_i \phi(x_i)$ and the dual problem:

$$\max_{\alpha \in \mathbb{R}^n} -g(-\lambda \alpha) - \frac{\lambda}{2} \alpha^\top K \alpha,$$

where, for $z \in \mathbb{R}^n$, $g(z) = \max_{u \in \mathbb{R}^n} -\frac{1}{n} \sum_{i=1}^{n} \ell(y_i, u_i) + u_i z_i$ is the Fenchel-conjugate of the empirical risk.

B Comparison of relative errors of kernel approximation and prediction performance

In Figure 3, we consider a prediction problem with $n = 400$ and a decay of eigenvalues of the kernel matrix which is the inverse of a low-order polynomial. We compare the decays to zero of the relative kernel matrix approximation $\|K - L\| / \|K\|$ (for the trace and operator norms) with the decay of the relative prediction performance (i.e., prediction for L minus prediction for the full matrix K).

As the rank p increases, the decay of the relative prediction error is much faster than the error in matrix approximation, suggesting that relying on good kernel matrix approximation may be suboptimal if the goal is simply to predict well.

C Proof of Theorem 1

We first prove a lemma that provides a Bernstein-type inequality for subsampled covariance matrices.

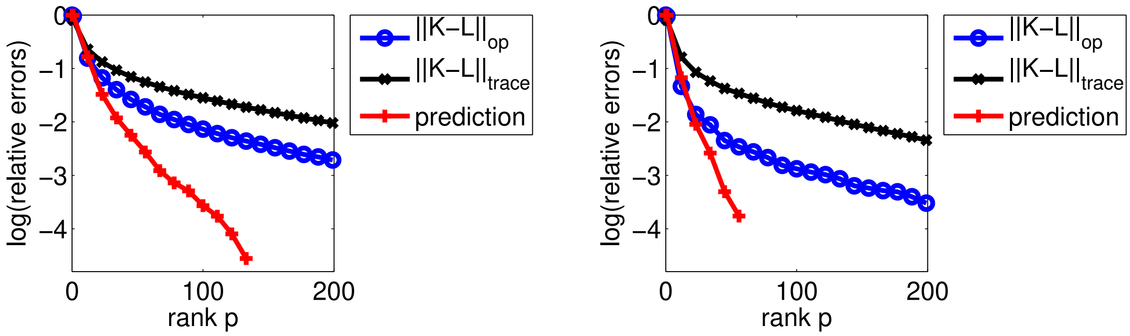

Figure 3: Relative average errors (in log-scale) as the rank p grows, for kernel matrix approximation $\|K - L\|$ (trace and operator norms) and predictions errors, for a synthetic prediction problem described in the experiments section of the main paper for $n = 400$. The prediction error (red curve) stops when the average prediction error of the column sampling approach gets below the prediction error of the full kernel matrix approach.

C.1 Concentration of subsampled covariance matrices

Given the matrix $\Psi \in \mathbb{R}^{n\times r}$ and $I \subset \{1,\dots,p\}$, we denote by Ψ_I the submatrix of Ψ composed of the rows of Ψ indexed by I.

Lemma 1 (Concentration of subsampled covariance) *Let $\Psi \in \mathbb{R}^{n\times r}$, with all rows of ℓ_2-norm less than R. Let I a random subset of $\{1,\dots,n\}$ with p elements (i.e., p elements chosen* without *replacement uniformly at random). Then, for all $t > 0$,*

$$\P_I\left(\lambda_{\max}\left[\frac{1}{n}\Psi^\top\Psi - \frac{1}{p}\Psi_I^\top\Psi_I\right] > t\right) \leqslant r\exp\left(\frac{-pt^2/2}{\lambda_{\max}(\frac{1}{n}\Psi^\top\Psi)(R^2+t/3)}\right).$$

Proof Let $\psi_1,\dots,\psi_n \in \mathbb{R}^r$ be the n rows of Ψ. We consider the matrix $\Delta \in \mathbb{R}^{r\times r}$ defined as:

$$\Delta = \frac{1}{n}\Psi^\top\Psi - \frac{1}{p}\Psi_I^\top\Psi_I = \frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^\top - \frac{1}{p}\sum_{i\in I}\psi_i\psi_i^\top.$$

By construction, we have $\mathbb{E}\Delta = 0$, and, as shown in [31, 32] and [33], we have

$$\mathbb{E}\,\mathrm{tr}\exp(s\Delta) \leqslant \mathbb{E}\,\mathrm{tr}\exp(s\Xi),$$

where Ξ is obtained by sampling independently p rows *with* replacement, i.e., is equal to

$$\Xi = \frac{1}{n}\sum_{i=1}^n \psi_i\psi_i^\top - \frac{1}{p}\sum_{j=1}^p\sum_{i=1}^n z_i^j\psi_i\psi_i^\top,$$

where $z^j \in \mathbb{R}^n$ is a random element of the canonical basis of $\mathbb{R}^n$ such that $\P(z_i^j = 1) = \frac{1}{n}$ for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,p\}$. This result extends to the matrix case the classical result of Hoeffding [34].

We thus have:

$$\Xi = \frac{1}{p}\sum_{j=1}^{p}\bigg(\frac{1}{n}\sum_{i=1}^{n}\psi_i\psi_i^\top - \sum_{i=1}^{n} z_i^j \psi_i\psi_i^\top\bigg) = \sum_{j=1}^{p} M_j,$$

with $M_j = \frac{1}{p}\big(\sum_{i=1}^n z_i^j\big(\frac{1}{n}\Psi^\top\Psi - \psi_i\psi_i^\top\big)\big)$. We have $\mathbb{E}M_j = 0$, $\lambda_{\max}(M_j) \leqslant \lambda_{\max}(\frac{1}{n}\Psi^\top\Psi)/p$, and

$$\begin{aligned}
\lambda_{\max}\bigg(\sum_{j=1}^{p}\mathbb{E}M_j^2\bigg) &= \frac{1}{p^2}\lambda_{\max}\bigg(\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{k=1}^{n}\mathbb{E}z_i^j z_k^j\big(\frac{1}{n}\Psi^\top\Psi - \psi_i\psi_i^\top\big)\big(\frac{1}{n}\Psi^\top\Psi - \psi_k\psi_k^\top\big)\bigg) \\
&= \frac{1}{p}\lambda_{\max}\bigg(\frac{1}{n}\sum_{i=1}^{n}\big(\frac{1}{n}\Psi^\top\Psi - \psi_i\psi_i^\top\big)^2\bigg) \text{ because } \mathbb{E}z_i^j z_k^j = \frac{1}{n}\delta_{i=k} \\
&= \frac{1}{p}\lambda_{\max}\big(\frac{1}{n}\sum_{i=1}^{n}\psi_i\psi_i^\top\psi_i\psi_i^\top - \big(\frac{1}{n}\Psi^\top\Psi\big)^2\big) \leqslant \frac{1}{p}\lambda_{\max}\big(\frac{1}{n}\sum_{i=1}^{n}\psi_i\psi_i^\top\psi_i\psi_i^\top\big) \\
&\leqslant \frac{R^2}{p}\lambda_{\max}\big(\frac{1}{n}\Psi^\top\Psi\big) \text{ because } \psi_i\psi_i^\top\psi_i\psi_i^\top \preccurlyeq R^2\psi_i\psi_i^\top.
\end{aligned}$$

We can then apply the matrix Bernstein inequality of [32, Theorem 6.1] to obtain the bound:

$$r\exp\bigg(-\frac{t^2/2}{\frac{R^2}{p}\lambda_{\max}\big(\frac{1}{n}\Psi^\top\Psi\big) + \frac{1}{p}\lambda_{\max}\big(\frac{1}{n}\Psi^\top\Psi\big)\frac{t}{3}}\bigg),$$

which leads to the desired result.

■

C.2 Proof of Theorem 1

Proof principle. Let $\Phi \in \mathbb{R}^{n\times n}$ be such that $K = \Phi\Phi^\top$. Note that if K has rank r, we could instead choose $\Phi \in \mathbb{R}^{n\times r}$.

We consider the regularized low-rank approximation $L_\gamma = \Phi N_\gamma \Phi^\top$, with

$$N_\gamma = \Phi_I^\top(\Phi_I\Phi_I^\top + p\gamma I)^{-1}\Phi_I = \Phi_I^\top\Phi_I(\Phi_I^\top\Phi_I + p\gamma I)^{-1} = I - \gamma(\Phi_I^\top\Phi_I/p + \gamma I)^{-1} \quad (8)$$

(using the matrix inversion lemma). We have $L = L_0$ but we will consider L_γ for $\gamma > 0$ to obtain a bound for $\gamma = 0$, using a monotonicity argument.

Following the same reasoning than in Section 4.1 of the main paper, the in-sample prediction error $\frac{1}{n}\mathbb{E}_\varepsilon\|\hat{z}_{L_\gamma} - z\|^2$ is equal to

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_\varepsilon\|\hat{z}_{L_\gamma} - z\|^2 &= n\lambda^2\|(\Phi N_\gamma\Phi^\top + n\lambda I)^{-1}z\|^2 + \frac{1}{n}\operatorname{tr} C\big[\Phi N_\gamma\Phi^\top(\Phi N_\gamma\Phi^\top + n\lambda I)^{-1}\big]^2 \\
&= \operatorname{bias}(L_\gamma) + \operatorname{variance}(L_\gamma).
\end{aligned}$$

The function $\gamma \mapsto N_\gamma$ is matrix-non-increasing (i.e., if $\gamma \geqslant \gamma'$, then $N_\gamma \preccurlyeq N_{\gamma'}$). Therefore, we have $0 \preccurlyeq N_\gamma \preccurlyeq N_0 \preccurlyeq I$. Since the variance term $\operatorname{variance}(L_\gamma) = \frac{1}{n}\operatorname{tr} C\big[\Phi N_\gamma\Phi^\top(\Phi N_\gamma\Phi^\top + n\lambda I)^{-1}\big]^2$

is non-decreasing in N_γ, this implies that the variance term with N_γ is smaller than the one with N_0 and then less then the one with N_γ replaced by I (which corresponds to the variance term without any approximation). For the bias term we have:

$$\text{bias}(L_\gamma) = n\lambda^2 \|(\Phi N_\gamma \Phi^\top + n\lambda I)^{-1} z\|^2 = n\lambda^2 z^\top (\Phi N_\gamma \Phi^\top + n\lambda I)^{-2} z, \tag{9}$$

which is a non-decreasing function of γ. Therefore, if we prove an upper-bound on the bias term for any $\gamma > 0$, we have a bound for $\gamma = 0$. This requires *lower-bounding* N_γ.

Lower-bounding N_γ. Let $\Psi = \Phi(\frac{1}{n}\Phi^\top\Phi + \gamma I)^{-1/2} \in \mathbb{R}^{n\times n}$. We may rewrite N_γ defined in Eq. (8) as

$$\begin{aligned}
N_\gamma &= I - \gamma(\frac{1}{p}\Phi_I^\top \Phi_I + \gamma I)^{-1} \\
&= I - \gamma\big(\frac{1}{n}\Phi^\top\Phi + \gamma I - \frac{1}{n}\Phi^\top\Phi + \frac{1}{p}\Phi_I^\top\Phi_I\big)^{-1} \\
&= I - \gamma(\frac{1}{n}\Phi^\top\Phi + \gamma I)^{-1/2}\Big[I - \frac{1}{n}\Psi^\top\Psi + \frac{1}{p}\Psi_I^\top\Psi_I\Big]^{-1}(\frac{1}{n}\Phi^\top\Phi + \gamma I)^{-1/2}.
\end{aligned}$$

Thus, in order to obtain a lower-bound on N_γ, it suffices to have an upper-bound of the form

$$\lambda_{\max}\big(\frac{1}{n}\Psi^\top\Psi - \frac{1}{p}\Psi_I^\top\Psi_I\big) \leqslant t, \tag{10}$$

which would imply

$$\begin{aligned}
I - N_\gamma &\preccurlyeq \frac{\gamma}{1-t}\big(\frac{1}{n}\Phi^\top\Phi + \gamma I\big)^{-1}, \\
K - L_\gamma &= \Phi(I - N_\gamma)\Phi^\top \preccurlyeq \frac{\gamma}{1-t}\Phi(\frac{1}{n}\Phi^\top\Phi + \gamma I)^{-1}\Phi^\top = \frac{n\gamma}{1-t}(K + n\gamma I)^{-1}K \preccurlyeq \frac{n\gamma}{1-t} I.
\end{aligned}$$

Assume $\frac{\gamma/\lambda}{1-t} \leqslant 1$. We then have, using the previous inequality:

$$(L_\gamma + n\lambda I)^{-1} \preccurlyeq \big(K - \frac{n\gamma}{1-t}I + n\lambda I\big)^{-1} = \big(K + n\lambda\big[1 - \frac{\gamma/\lambda}{1-t}\big]I\big)^{-1} \preccurlyeq (1 - \frac{\gamma/\lambda}{1-t})^{-1}(K + n\lambda I)^{-1}.$$

Thus, the bias term in Eq. (9) is less than the original bias term times $(1 - \frac{\gamma/\lambda}{1-t})^{-2}$. If the bound defined in Eq. (10) is not met, then we can upper-bound the bias term by $\frac{1}{n}z^\top z$, which is itself upper-bounded by the unapproximated bias term times $(1+\frac{R^2}{\lambda})$—indeed, we have $n\lambda^2 z^\top (K + n\lambda I)^{-2} z \geqslant n\lambda^2 z^\top z(n\lambda + nR^2)^{-2} = \frac{1}{n}z^\top z(1 + R^2/\lambda)^{-2}$. Thus if we define $p_t = ¶_I\Big(\lambda_{\max}\Big[\frac{1}{n}\Psi^\top\Psi - \frac{1}{p}\Psi_I^\top\Psi_I\Big] > t\Big)$, then, $E_I\big[\text{bias}(L_\gamma)\big]$ is upper-bounded by

$$B = p_t(1 + R^2/\lambda) + (1 - p_t)\big(1 - \frac{\gamma/\lambda}{1-t}\big)^{-2}, \tag{11}$$

times $\text{bias}(K)$.

Probabilistic control. We need to upper-bound the largest eigenvalue of $\frac{1}{n}\Psi^\top\Psi - \frac{1}{p}\Psi_I^\top\Psi_I$, where I is a random subset of $\{1,\dots,n\}$ of cardinality p. This is the difference between an empirical second-order moment and the empirical moment of a subset of p random elements. In order to apply Lemma 1, we need to upper-bound the squared ℓ_2-norm as (assuming $\gamma \leqslant \lambda$):

$$\begin{aligned}
\max_{i\in\{1,\dots,n\}} \big(\Psi\Psi^\top\big)_{ii} &= \max_{i\in\{1,\dots,n\}} \big(\Phi(\frac{1}{n}\Phi^\top\Phi + \gamma I)^{-1}\Phi^\top\big)_{ii} \\
&= \lambda\gamma^{-1} \max_{i\in\{1,\dots,n\}} \big(\Phi(\frac{1}{n}(\lambda\gamma^{-1})\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top\big)_{ii} \\
&\leqslant \lambda\gamma^{-1} \max_{i\in\{1,\dots,n\}} \big(\Phi(\frac{1}{n}\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top\big)_{ii} \text{ because } \gamma\leqslant\lambda, \\
&= n\lambda\gamma^{-1}\big\| \operatorname{diag}\big(K(K+n\lambda I)^{-1}\big)\big\|_\infty = \lambda\gamma^{-1}d.
\end{aligned}$$

Thus for $\gamma \leqslant \lambda$, all rows of Ψ have a squared ℓ_2-norm upper-bounded by $\lambda\gamma^{-1}d$, and $\frac{1}{n}\Psi^\top\Psi \preccurlyeq I$, we can apply Lemma 1, to obtain that:

$$p_t = ¶_I\Big(\lambda_{\max}\Big[\frac{1}{n}\Psi^\top\Psi - \frac{1}{p}\Psi_I^\top\Psi_I\Big] > t\Big) \leqslant n\exp\left(\frac{-pt^2/2}{\lambda\gamma^{-1}d + t/3}\right).$$

Using the bound from Eq. (11), we get, given $\delta\in(0,1)$, $t=1/2$, and $\gamma=\frac{\lambda\delta}{4}$,

$$\begin{aligned}
B &= 1 + \frac{R^2}{\lambda}p_t + (1-p_t)\Big[\big(1-\frac{\gamma/\lambda}{1-t}\big)^{-2} - 1\Big] \\
&\leqslant 1 + \frac{nR^2}{\lambda}\exp\left(\frac{-p/8}{4d/\delta + 1/6}\right) + \Big[(1-\delta/2)^{-2} - 1\Big] \\
&\leqslant 1 + \frac{nR^2}{\lambda}\exp\left(\frac{-p}{32d/\delta + 2}\right) + \Big[\frac{\delta - \delta^2/4}{(1-\delta/2)^2}\Big] \\
&\leqslant 1 + \frac{nR^2}{\lambda}\exp\left(\frac{-p}{32d/\delta + 2}\right) + \delta\Big[\frac{1-\delta/4}{(1-\delta/2)^2}\Big] \\
&\leqslant 1 + \frac{nR^2}{\lambda}\exp\left(\frac{-p}{32d/\delta + 2}\right) + \delta\Big[\frac{3/4}{1/4}\Big] = \frac{nR^2}{\lambda}\exp\left(\frac{-p}{32d/\delta + 2}\right) + 3\delta.
\end{aligned}$$

Thus, if $p \geqslant \big(\frac{32d}{\delta} + 2\big)\log\frac{nR^2}{\delta\lambda}$, we obtain that $B \leqslant 1 + 4\delta$.

Thus,

$$\begin{aligned}
\mathbb{E}_I\big[\text{bias}(L) + \text{variance}(L)\big] &\leqslant \mathbb{E}_I\big[\text{bias}(L_\gamma) + \text{variance}(K)\big] \text{ by monotonicity} \\
&= \mathbb{E}_I\big[\text{bias}(L_\gamma)\big] + \text{variance}(K) \\
&\leqslant (1+4\delta)\text{bias}(K) + \text{variance}(K) \\
&\leqslant (1+4\delta)\big[\text{bias}(K) + \text{variance}(K)\big],
\end{aligned}$$

which is the desired result. Note that

- We could improve the bound by expliciting the reduction of the variance term.
- In some situations, the prediction performance for the approximated version may in fact be smaller than the non-approximated version.

D Asymptotics of bias and variance terms

In this appendix, we consider various decays of eigenvalues $n\mu_i$ of K and components $\sqrt{n\nu_i}$ (in magnitude) of the signal z to estimate. We follow the reasoning of [35] (i.e., replacing sums by integrals). Given our assumptions, we have:

$$\begin{aligned}
\text{bias} &= n^2\lambda^2 \sum_{i=1}^{n} \frac{n\nu_i}{(n\mu_i + n\lambda)^2} = n\lambda^2 \sum_{i=1}^{n} \frac{\nu_i}{(\mu_i+\lambda)^2}, \\
\frac{n}{\sigma^2}\text{variance} &= \sum_{i=1}^{n} \frac{n^2\mu_i^2}{(n\mu_i+n\lambda)^2} = \sum_{i=1}^{n} \frac{\mu_i^2}{(\mu_i+\lambda)^2}.
\end{aligned}$$

For all cases we need to consider, for simplicity, we only provide an upper-bound for μ_i exactly equal to its asymptotic equivalent. Considering lower-bounds and a constant times the asymptotic equivalent may be done in a similar way.

D.1 Variance terms

We consider the two possible cases (the variance term only depends on (μ_i)). Moreover we show that the two traditional definitions of the degrees of freedom, $\operatorname{tr} K(K+n\lambda I)^{-1}$ and $\operatorname{tr} K^2(K+n\lambda I)^{-2}$, have the same asymptotically equivalents.

Polynomial decay ($\mu_i = i^{-2\beta}$, $\beta > 1/2$). The renormalized variance term is less than

$$\begin{aligned}
\sum_{i=1}^{n} \frac{1}{(1+i^{2\beta}\lambda)^2} &\leqslant \int_0^n \frac{1}{(1+t^{2\beta}\lambda)^2} dt \\
&= \int_0^{\lambda n^{2\beta}} \frac{1}{(1+u)^2} \lambda^{-1/2\beta} u^{1/2\beta - 1} \frac{1}{2\beta} du \text{ with the change of variable } u = \lambda t^{2\beta}, \\
&\leqslant \int_0^{\infty} \frac{1}{(1+u)^2} \lambda^{-1/2\beta} u^{1/2\beta - 1} \frac{1}{2\beta} du \\
&= O(\lambda^{-1/2\beta}) \text{ since the integral is finite.}
\end{aligned}$$

With the same reasoning, we have $\operatorname{tr} K(K+n\lambda I)^{-1} \leqslant \int_0^{\lambda n^{2\beta}} \frac{1}{(1+u)} \lambda^{-1/2\beta} u^{1/2\beta-1} \frac{1}{2\beta} du = O(\lambda^{-1/2\beta})$.

Exponential decay ($\mu_i = e^{-\rho i}$). The renormalized variance term is less than

$$\begin{aligned}
\sum_{i=1}^n \frac{1}{(1+e^{\rho i}\lambda)^2} &\leqslant \int_0^n \frac{1}{(1+e^{\rho t}\lambda)^2}dt = \int_0^n \frac{e^{-2\rho t}}{(e^{-\rho t}+\lambda)^2}dt \\
&= \frac{1}{\rho}\int_{e^{-\rho n}}^1 \frac{u}{(u+\lambda)^2}du \text{ with the change of variable } u = e^{-\rho t} \\
&\leqslant \frac{1}{\rho}\int_0^1 \Big(\frac{1}{u+\lambda} - \frac{\lambda}{(u+\lambda)^2}\Big)du \leqslant \frac{1}{\rho}\int_0^1 \frac{1}{u+\lambda}du \\
&= \frac{1}{\rho}\big[\log(1+\lambda) - \log\lambda\big] = O(\log\frac{1}{\lambda}).
\end{aligned}$$

We the same technique, we get bounds on $\operatorname{tr} K(K+n\lambda I)^{-1}$ in the same way we just did for $\operatorname{tr} K^2(K+n\lambda I)^{-2}$.

D.2 Bias terms

The bias terms depend on both (μ_i) and (ν_i) and we consider all combinations.

Polynomial decays ($\mu_i = i^{-2\beta}$, $\nu_i = i^{-2\delta}$, $\beta, \delta > 1/2$). The bias term is less than

$$n\lambda^2 \sum_{i=1}^n \frac{i^{4\beta-2\delta}}{(1+i^{2\beta}\lambda)^2} \leqslant 2n\lambda^2 \int_1^n \frac{t^{4\beta-2\delta}}{(1+t^{2\beta}\lambda)^2}dt. \tag{12}$$

If $2\delta - 4\beta > 1$, then we have an upper bound of $2n\lambda^2\int_1^\infty t^{4\beta-2\delta}dt = O(n\lambda^2)$, because the integral is finite.

If $2\delta - 4\beta < 1$, then we can further bound Eq. (12) as

$$\begin{aligned}
2n\lambda^2\int_1^n \frac{t^{4\beta-2\delta}}{(1+t^{2\beta}\lambda)^2}dt &= 2n\lambda^2\int_\lambda^{n^{2\beta}\lambda} \frac{u^{2-\delta/\beta+\frac{1}{2\beta}-1}\lambda^{-2+\delta/\beta-\frac{1}{2\beta}}}{(1+u)^2}\frac{1}{2\beta}du \\
&\quad \text{with the change of variable } u = \lambda t^{2\beta} \\
&= O(\lambda^{(2\delta-1)/2\beta})\int_0^\infty \frac{u^{2-\delta/\beta+\frac{1}{2\beta}-1}}{(1+u)^2}\frac{1}{2\beta}du = O(\lambda^{(2\delta-1)/2\beta}),
\end{aligned}$$

because the integral is finite (due to the assumptions made on β and δ).

Exponential decays ($\mu_i = e^{-\rho i}$, $\nu_i = e^{-\kappa i}$, $\rho, \kappa > 0$). The bias term is less than

$$\begin{aligned}
n\lambda^2\sum_{i=1}^n \frac{e^{(2\rho-\kappa)i}}{(1+e^{\rho i}\lambda)^2} &\leqslant n\lambda^2\int_1^n \frac{e^{(\rho-\kappa)t}}{(1+e^{\rho t}\lambda)^2}e^{\rho t}dt \\
&= \frac{n\lambda}{\rho}\int_\lambda^{\lambda e^{n\rho}} \frac{(u/\lambda)^{1-\kappa/\rho}}{(1+u)^2}du \text{ with the change of variables } u = \lambda e^{\rho t}.
\end{aligned}$$

If $\kappa/\rho > 2$, then we have a bound

$$\frac{n\lambda^2}{\rho}\int_1^\infty \frac{u^{1-\kappa/\rho}}{(1+\lambda u)^2}du = O(n\lambda^2),$$

because the integral is finite and uniformly bounded in λ.

If $\kappa/\rho < 2$, then we have a bound

$$\frac{n\lambda^{\kappa/\rho}}{\rho}\int_\lambda^{\lambda e^{n\rho}} \frac{u^{1-\kappa/\rho}}{(1+u)^2}du \leqslant \frac{n\lambda^{\kappa/\rho}}{\rho}\int_0^\infty \frac{u^{1-\kappa/\rho}}{(1+u)^2}du = O(n\lambda^{\kappa/\rho}).$$

Mixed decays. For μ_i with polynomial decays and ν_i with exponential decays, we are in a situation where ν_i is decaying fast enough (faster than $i^{-2\delta}$ for any $\delta > 1/2$) so that, given previous results, the bias is $n\lambda^2$.

The only remaining result to show is $\mu_i = e^{-\rho i}$ and $\nu_i = i^{-2\delta}$, $\delta > 1/2$, which we now consider. The bias term is equal to

$$\begin{aligned}
n\lambda^2\sum_{i=1}^n \frac{\nu_i}{(\mu_i+\lambda)^2} &= n\lambda^2\sum_{i=1}^n \frac{i^{-2\delta}}{(e^{-\rho i}+\lambda)^2} \\
&= n\lambda^2 \sum_{i\leqslant \frac{1}{\rho}\log\lambda^{-1}} \frac{i^{-2\delta}}{(e^{-\rho i}+\lambda)^2} + n\lambda^2 \sum_{n\geqslant i>\frac{1}{\rho}\log\lambda^{-1}} \frac{i^{-2\delta}}{(e^{-\rho i}+\lambda)^2} \\
&\leqslant n\lambda^2 \sum_{i\leqslant \frac{1}{\rho}\log\lambda^{-1}} \frac{i^{-2\delta}}{e^{-2\rho i}} + n\lambda^2 \sum_{n\geqslant i>\frac{1}{\rho}\log\lambda^{-1}} \frac{i^{-2\delta}}{\lambda^2} \\
&\leqslant n \sum_{i\leqslant \frac{1}{\rho}\log\lambda^{-1}} i^{-2\delta} + n \sum_{i>\frac{1}{\rho}\log\lambda^{-1}} i^{-2\delta} \\
&= O(n) + O(n(\log\lambda^{-1})^{1-2\delta}) = O(n(\log\lambda^{-1})^{1-2\delta}).
\end{aligned}$$

D.3 Optimal regularization parameters

We can now take all six cases, and compute the optimal λ and the resulting optimal regularization error.

- $\mu_i = i^{-2\beta}$, $\nu_i = i^{-2\delta}$ ($2\delta > 4\beta+1$): we need to minimize with respect to λ the function $n^{-1}\lambda^{-1/2\beta} + \lambda^2$, which leads to $\lambda \approx n^{-1/(2+1/2\beta)}$ and an optimal value of $n^{1/(4\beta+1)-1}$.
- $\mu_i = i^{-2\beta}$, $\nu_i = i^{-2\delta}$ ($2\delta < 4\beta+1$): we need to minimize with respect to λ the function $n^{-1}\lambda^{-1/2\beta} + \lambda^{(2\delta-1)/2\beta}$, which leads to $\lambda \approx n^{-\beta/\delta}$ and an optimal value of $n^{1/(2\delta)-1}$.
- $\mu_i = i^{-2\beta}$, $\nu_i = e^{-\kappa i}$: same computation as the first one.
- $\mu_i = e^{-\rho i}$, $\nu_i = i^{-2\delta}$: we need to minimize with respect to λ the function $n^{-1}\log\frac{1}{\lambda} + (\log\frac{1}{\lambda})^{1-2\delta}$, which leads to $\log\frac{1}{\lambda} \approx n^{1/2\delta}$ and an optimal value of $n^{1/2\delta-1}$.

– $\mu_i = e^{-\rho i}$, $\nu_i = e^{-\kappa i}$ ($\kappa > 2\rho$): we need to minimize with respect to λ the function $n^{-1} \log \frac{1}{\lambda} + \lambda^2$, which leads to $\lambda \approx n^{-1/2}$ and an optimal value of $\log n/n$.

– $\mu_i = e^{-\rho i}$, $\nu_i = e^{-\kappa i}$ ($\kappa < 2\rho$): we need to minimize with respect to λ the function $n^{-1} \log \frac{1}{\lambda} + \lambda^{\kappa/\rho}$, which leads to $\lambda \approx n^{-\rho/\kappa}$ and an optimal value of $\log n/n$.

E Kernels on [0,1]

In this appendix, we consider kernels on $\mathcal{X} = [0, 1]$ that lead to closed-form expressions (or asymptotic equivalents) for eigenvalues of K and components of z. These are used in simulations.

Kernels. For a positive summable sequence $(\mu_i)_{i \geqslant 1}$, we consider $k(x, y) = \sum_{i=1}^{\infty} 2\mu_i \cos 2i\pi(x-y)$. It is defined for any $(x, y) \in [0, 1]^2$ and is 1-periodic in x and y. It is a function g of $x - y - \lfloor x - y \rfloor$, i.e., $k(x, y) = g(x - y - \lfloor x - y \rfloor)$. Moreover $k(x, x)$ is independent of x.

For $\mu_i = \frac{1}{i^{2\beta}}$, we have $k(x, y) = \frac{1}{(2\beta)!} B_{2\beta}(x - y - \lfloor x - y \rfloor)$, where $B_{2\beta}$ is the (2β)-th Bernoulli polynomial [24]. For example, we have $B_2(x) = x^2 - x + \frac{1}{6}$ and $B_6(x) = x^6 - 3x^5 + \frac{5}{2}x^4 - \frac{1}{2}x^2 + \frac{1}{42}$.

For $\mu_i = e^{-\rho i}$, we have, $k(x, y) = 2\frac{e^{\rho} \cos 2\pi(x-y) - 1}{e^{2\rho} - 2e^{\rho} \cos 2\pi(x-y) + 1}$. Indeed, we have

$$\begin{aligned}
k(x, y) &= \mathrm{Re}\bigg(\sum_{i=1}^{\infty} 2e^{-\rho i + 2i\omega\pi(x-y)} \bigg) \text{ with } \omega^2 = -1, \\
&= 2\mathrm{Re}\bigg(\sum_{i=1}^{\infty} e^{i(-\rho + 2\omega\pi(x-y))} \bigg) = 2\mathrm{Re}\bigg(\frac{e^{-\rho + 2\omega\pi(x-y)}}{1 - e^{-\rho + 2\omega\pi(x-y)}} \bigg) \\
&= 2\mathrm{Re}\bigg(\frac{1}{e^{\rho - 2\omega\pi(x-y)} - 1} \bigg) = 2\frac{e^{\rho} \cos 2\pi(x - y) - 1}{e^{2\rho} - 2e^{\rho} \cos 2\pi(x - y) + 1}.
\end{aligned}$$

Data and eigenvectors. If we consider n data points $x_i = \frac{i-1}{n}$, $i = 1, \dots, n$, then the kernel matrix K has components $K_{ij} = k(\frac{i-1}{n}, \frac{j-1}{n})$. It is a circulant matrix, thus it is diagonalisable in the discrete Fourier basis [36], with eigenvalues equal to the discrete Fourier transform of the first column of the matrix, i.e., $(g(0), g(1/n), \dots, g(1 - 1/n))^{\top}$.

Thus, the i-th eigenvector has j-th component $\frac{1}{\sqrt{n}}e^{2\omega i(j-1)\pi/n}$ (with $\omega^2 = 01$) and the i-th eigenvalue is

$$\begin{aligned}
\lambda_i &= \sum_{j=1}^{n} e^{-2\omega i(j-1)\pi/n} g((j-1)/n) \\
&= \sum_{j=1}^{n} e^{-2\omega i(j-1)\pi/n} \sum_{s=1}^{\infty} 2\mu_s \cos 2s\pi(j-1)/n \\
&= \sum_{j=1}^{n} e^{-2\omega i(j-1)\pi/n} \sum_{s=1}^{\infty} \mu_s \big[e^{2s\omega\pi(j-1)/n} + e^{-2s\omega\pi(j-1)/n}\big] \\
&= n \sum_{s=1}^{\infty} \mu_s \big[\delta_{s=i[n]} + \delta_{-s=i[n]}\big] \text{ because of the orthonormality of the Fourier basis,} \\
&= n\mu_i + n\sum_{h=1}^{\infty} \mu_{i+hn} + n\sum_{h=1}^{\infty} \mu_{-i+hn}.
\end{aligned}$$

If n is large and μ_i tends to zero when i tends to $+\infty$, then an asymptotic equivalent for λ_i is $n\mu_i$.

For data sampled from the uniform distribution in $[0, 1]$, then similar equivalents hold (see, e.g., [35]).

Functions. Let $f(x) = \sum_{i=1}^{\infty} 2\nu_i^{1/2} \cos 2i\pi x$, for ν_i a non-negative summable sequence. We consider $z_i = f(x_i) = f((i-1)/n)$. The component of z on the i-th eigenvector of K is (following the same reasoning as above):

$$\begin{aligned}
& \sum_{j=1}^{n} \frac{1}{\sqrt{n}} e^{-2\omega i(j-1)\pi/n} f((j-1)/n) \\
= \;& \sqrt{n}\bigg(\nu_i^{1/2} + \sum_{h=1}^{\infty} \nu_{i+hn}^{1/2} + \sum_{h=1}^{\infty} \nu_{-i+hn}^{1/2}\bigg),
\end{aligned}$$

and the asymptotic equivalent is $(n\nu_i)^{1/2}$.

Link with Sobolev spaces. The kernel $k(x, y)$ defined above corresponds for $\mu_i = i^{-2\beta}$ to certain Sobolev spaces [24, 26]. Indeed, for β integer, the associated RKHS is the Sobolev space of periodic functions which are β-times differentiable.

Moreover, when $\nu_i = i^{-2\delta}$, then for $\delta > \delta_0$, then the corresponding function is $(\delta_0 - 1/2)$-times differentiable, and the minimax rate of estimation is known to be exactly $O(n^{1/2\delta_0})$ [37, 29]. Thus, up to logarithmic terms, the best possible rate is $O(n^{1/2\delta})$, and is achieved if β is large enough (see Section 4.3 of the main paper).

References

[1] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[2] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Camb. U. P., 2004.

[3] B. Schölkopf, K. Tsuda, and J.P. Vert. *Kernel methods in computational biology*. MIT press, 2004.

[4] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. In *CVPRW*, 2006.

[5] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[6] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Adv. NIPS*, 2001.

[7] A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proc. ICML*, 2000.

[8] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *J. Mac. Learn. Res.*, 2:243–264, 2001.

[9] M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

[10] C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proc. SODA*, 2009.

[11] A. Talwalkar S. Kumar, M. Mohri. Sampling methods for the Nyström method. *JMLR*, 13:981–1006, 2012.

[12] C. Cortes, M. Mohri, and A. Talwalkar. On the impact of kernel approximation on learning accuracy. In *Proc. AISTATS*, 2010.

[13] R. Jin, T. Yang, M. Mahdavi, Y.-F. Li, and Z.-H. Zhou. Improved bound for the Nyström's method and its application to kernel classification. Technical Report 1111.2262v2, arXiv, 2001.

[14] O. Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178, 2007.

[15] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proc. ICML*, 2007.

[16] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59, 2009.

[17] J.C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*, pages 185–208. MIT Press, 1999.

[18] O. Dekel, S. Shalev-Shwartz, and Y. Singer. The Forgetron: A kernel-based perceptron on a fixed budget. In *Adv. NIPS*, 2005.

[19] F. Orabona, J. Keshet, and B. Caputo. The Projectron: a bounded kernel-based perceptron. In *Proc. ICML*, 2008.

[20] S. S. Keerthi, O. Chapelle, and D. DeCoste. Building support vector machines with reduced classifier complexity. *JMLR*, 7:1493–1515, 2006.

[21] N. D. Lawrence, M. Seeger, and R. Herbrich. Fast sparse Gaussian process methods: The informative vector machine. In *Adv. NIPS*, 2002.

[22] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

[23] O. A. Maillard and R. Munos. Compressed least-squares regression. In *Adv. NIPS*, 2009.

[24] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[25] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

[26] C. Gu. *Smoothing spline ANOVA models*. Springer, 2002.

[27] D. Hsu, S. M. Kakade, and T. Zhang. An analysis of random design linear regression. Technical Report 1106.2363, Arxiv, 2011.

[28] S. Arlot and F. Bach. Data-driven calibration of linear estimators with minimal penalties. In *Adv. NIPS*, 2009.

[29] I. M. Johnstone. Minimax Bayes, asymptotic minimax and sparse wavelet priors. *Statistical Decision Theory and Related Topics*, pages 303–326, 1994.

[30] F. Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[31] J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Arxiv preprint arXiv:1011.1595*, 2010.

[32] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, pages 1–46, 2011.

[33] A. Gittens. The spectral norm error of the naive Nyström extension. *Arxiv preprint arXiv:1110.5305*, 2011.

[34] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[35] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis, April 2008. oai:hal.archives-ouvertes.fr:hal-00270806_v1.

[36] R. M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

[37] P. Speckman. Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics*, 13(3):970–983, 1985.