Reducing training time by efficient localized kernel regression

Nicole Mücke
University of Stuttgart
nicole.muecke@mathematik.uni-stuttgart.de

Abstract

We study generalization properties of kernel regularized least squares regression based on a partitioning approach. We show that optimal rates of convergence are preserved if the number of local sets grows sufficiently slowly with the sample size. Moreover, the partitioning approach can be efficiently combined with local Nyström subsampling, improving computational cost twofold.

1 Introduction

The use of reproducing kernel methods for nonparametric regression such as *Kernel Regularized Least Squares* (KRLS) or the Support Vector Machine has enjoyed a wide popularity and their theoretical properties are well understood. These methods are attractive because they attain asymptotically minimax optimal rates of convergence. But it is also well known that they scale poorly when massive datasets are involved. Large training sets give rise to large computational and storage costs. For example, computing a kernel ridge regression estimate needs inversion of a $n \times n$-matrix, with n the sample size. This requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory, which becomes prohibitive for large sample sizes.

Large Scale Problems: Subsampling and Localization. Because of the above mentioned shortcomings various methods have been developed for saving computation time and memory requirements, speeding up the usual approaches. During the last years, a huge amount of research effort was devoted to finding *low-rank approximations* of the kernel matrix. A popular instance is Nyström sampling see e.g. Williams and Seeger (2000), Bach (2013), Rudi et al. (2015) where one aims at replacing the theoretically optimal approximation obtained by a spectral decomposition (which requires time at least $\mathcal{O}(n^2)$) by a less ambitious suitable low rank approximation of the kernel matrix via column sampling, reducing run time to $\mathcal{O}(np^2)$ where p denotes the rank of the approximation. Clearly the rules of the game are to choose p as small as possible while maintaining minimax optimality of convergence rates and to explicitly determine this p as a function of the sample size n.

Another line of research with computational benefits is devoted to so called *partition-based* or *localized* approaches, see Segata and Blanzieri (2010) for localized SVMs for binary classification, Meister and Steinwart (2016) for localized SVMs using the Gaussian RBF kernel or Tandon et al. (2016) for more general kernels in an KRLS framework. The main idea behind the partitioning approach is to split the training data based on a disjoint partition of the input space into smaller subsamples and to train only on smaller chunks. Prediction for a new input is then much faster since one only has to identify the local subset to which the new input belongs and to use the local estimator.

Another benefit in using localized approaches lies in exploiting regions of high regularity. It is well known that rates of convergence highly depend on regularity: The smoother the objective function, the faster the rate of convergence. The usual global learning approach however doesn't "see" regions of higher regularity. Global rates of convergence are determined by the region of the input space where the target is least smooth.

Our results show, when building an KRLS estimator based on accurate local ones trained on subregions of the training set, we better take into account the local regularity of the objective function, leading to more accurate local approximations. In particular, our approach does not suffer from local underfitting, even though the regularization parameter is chosen as in the global approach.

Further, we show that the partitioning approach for

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

KRLS can be efficiently combined with Nyström subsampling, substantially reducing training time and speeding up the more usual (localized) version of KRLS.

Informally, we show if the number of subsets is *not too large* and if the number of subsampled datapoints is *large enough* we obtain fast upper rates of convergence. An important aspect of our approach is the observation that under appropriate conditions on the probability of subsamples - which come quite naturally in the partitioning approach - our rates of convergence are actually guided by local regions of high regularity, leading to improved finite sample bounds.

In this paper, we shall focus only on KRLS, although our results could be extended to a much larger class of general spectral regularization methods, including e.g. Gradient Descent, similar to Rosasco et al. (2005) , Dicker et al. (2017) , Blanchard and Mücke (2017) or more recent Lin et al. (2018). For a more detailed discussion of our results and a comparison to related research we refer to Section 6.

The outline of our paper is as follows: Section 2 is devoted to an introduction to the learning problem in an RKHS framework. In Section 3 we firstly introduce the partitioning approach and introduce all assumptions needed to establish our main Theorems. In Section 4 we briefly recall the Nystöm method and give an upper bound in expectation for the rate of convergence. Section 5 is devoted to showing that the partitioning approach and subsampling can be effeciently combined. Finally, we compare our results with other approaches in Section 6 and finish with a conclusion in Section 7 . All our proofs a deferred to the Appendix.

Notation: For $n \in \mathbb{N}$, we denote by $[n]$ the set of integers $\{1, ..., n\}$. For two positive sequences $(a_n)_n$ and $(b_n)_n$, the expression $a_n \lesssim b_n$ means that $a_n \leq Cb_n$, for some universal constant $C < \infty$. For f in a Hilbert space $\mathcal{H}$ we let $f \otimes f$ be the outer product acting as rank-one operator $(f \otimes f)h = \langle h, f\rangle_{\mathcal{H}} f$.

2 Learning with Kernels

In this section we introduce the supervised learning problem and give an overview of regularized learning in an RKHS framework.

Learning Setting. We consider the well-established setting of learning under random design where $\mathcal{X} \times \mathbb{R}$ is a probability space with distribution ρ. We let ν be the marginal distribution on $\mathcal{X}$ and $\rho(\cdot|x)$ denotes the conditional distribution on $\mathbb{R}$ given $x \in \mathcal{X}$. Our goal is minimizing the *expected risk*

$$\mathcal{E}(f) = \int_{\mathcal{X}\times\mathbb{R}} (f(x) - y)^2 \, d\rho(x, y) \, .$$

It is known that this quantity is minimized over $L^2(\nu)$ by the regression function

$$f_\rho(x) = \int_{\mathcal{Y}} y \, d\rho(y|x) \, .$$

However, we exclusively focus our analysis to the special case where f_ρ lies in a hypothesis space $\mathcal{H} \subset L^2(\nu)$ of measurable functions from $\mathcal{X}$ to $\mathbb{R}$.

We are particularly interested in the case where $\mathcal{H}$ is a separable *reproducing kernel Hilbert space* (RKHS), possessing a bounded positive definite symmetric measurable kernel K on $\mathcal{X}$. Throughout the paper we assume that

Assumption 1.

$$\kappa^2 := \sup_{x,x'} K(x, x') < \infty \, .$$

An important feature is the *reproducing property:* For any $x \in \mathcal{X}$ and any $f \in \mathcal{H}$ one has

$$f(x) = \langle f, K_x\rangle_{\mathcal{H}} \, ,$$

where $K_x := K(x, \cdot) \in \mathcal{H}$, see e.g. Aronszajn (1950).

Given a sample $\mathbf{z} = \{z_j = (x_j, y_j)\}_{j=1}^n$ of size $n \in \mathbb{N}$, a classical approach for empirically solving the minimization problem described above is by *Kernel Regularized Least Squares* (KRLS), also known as *Tikhonov Regularization*. This approach is based on minimization of the penalized empirical functional

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{j=1}^{n}(f(x_j) - y_j)^2 + \lambda||f||_{\mathcal{H}}^2 \, , \tag{1}$$

where $\lambda > 0$ is the *regularization parameter*. The Representer Theorem, see e.g. Steinwart and Christmann (2008), ensures that the solution $\hat{f}_{\mathbf{z}}^{\lambda}$ to (1) exists, is unique and can be written as

$$\hat{f}_{\mathbf{z}}^{\lambda}(x) = \sum_{j=1}^{n} \alpha_j K(x_j, x) \tag{2}$$

with

$$\alpha = (\mathbb{K}_n + \lambda n I)^{-1}\mathbf{y}$$

and where $\mathbb{K}_n = (K(x_i, x_j))_{i,j} \in \mathbb{R}^{n\times n}$ is the kernel matrix. In particular, this means that minimization can be restricted to the space

$$\mathcal{H}_n = \{ \, f \in \mathcal{H} \mid f = \sum_{j}^{n} \alpha_j K(x_j, \cdot) \, , \alpha_j \in \mathbb{R} \, \} \, .$$

Rates of convergence and Optimality. A common goal of learning theory is to give upper bounds for the convergence of $\hat{f}_{\mathbf{z}}^{\lambda_n}$ to f_ρ, where the regularization parameter is tuned according to sample size, and derive rates of convergence as $n \to \infty$ under appropriate assumptions on the regularity of f_ρ. In this paper, our bounds are given in the usual squared $L^2(\nu)$ distance with respect to the sampling distribution, which is equal to the excess risk when using the squared loss, i.e.

$$\|\hat{f}^\lambda - f_\rho\|_{L^2}^2 = \mathcal{E}(f^\lambda) - \mathcal{E}(f_\rho) \,. \tag{3}$$

More precisely, we are interested in bounding the averaged above error over the draw of the training data (this is also called Mean Integrated Squared Error).

A common framework for expressing regularity of the target function is by means of the kernel covariance operator

$$T = \mathbb{E}[K_X \otimes K_X] \,.$$

If there exists $r > 0$ such that

$$||T^{-r} f_\rho||_{\mathcal{H}} \leq R \tag{4}$$

for some $R < \infty$, then f_ρ is considered as regular. In particular, this assumption ensures that $f_\rho \in \mathcal{H}$, see e.g. Fischer and Steinwart (2017). This type of regularity class, also called *source condition*, has been considered in a learning context by Cucker and Smale (2002), and Caponnetto and De Vito (2006) have established upper bounds for the performance of KRLS over such classes. This has been extended to other types of kernel regularization methods by Caponnetto and Yao (2010); Dicker et al. (2017); Blanchard and Mücke (2017).

Furthermore, bounds on the generalization error also depend on the notion of *effective dimension* of the data with respect to the regularization parameter λ, defined as

$$\mathcal{N}(\lambda) := \mathcal{N}(T, \lambda) := \mathrm{Trace}[(T + \lambda)^{-1} T] \,. \tag{5}$$

An assumed bound of the form

$$\mathcal{N}(\lambda) \lesssim \lambda^{-\gamma}$$

with $0 < \gamma \leq 1$ is referred to as a *Capacity Assumption*, see Zhang (2005). In particular, it is shown in Caponnetto and De Vito (2006) that (5) is ensured if the eigenvalues[1] $(\mu_j)_j$ of T enjoy a polynomial decay, i.e. $\mu_j \lesssim j^{-\frac{1}{\nu}}$.

It is well known and fairly standard that bounds of the excess risk (3) are guided by the two conditions (4) on regularity and (5) on the capacity, i.e.

$$\mathbb{E}\big[\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho)\big] \lesssim R^2 \left(\frac{1}{n}\right)^{\frac{2r+1}{2r+1+\gamma}} \,, \tag{6}$$

[1]Note that boundednes of K ensures that T is trace class, hence compact and has a discrete spectrum.

with $0 < r \leq \frac{1}{2}$, provided the regularization parameter is chosen according to

$$\lambda_n \simeq \left(\frac{1}{n}\right)^{\frac{1}{2r+1+\gamma}} \,. \tag{7}$$

In the framework of KRLS, these bounds were derived in Caponnetto and De Vito (2006); Blanchard and Mücke (2017) derive bounds in a more general framework. Both papers also show optimality (i.e. there is also a corresponding lower bound).

From (6) we immediately see that the regularity inherent in the problem has an impact on the speed of convergence: The larger the regularity, the faster is convergence.

3 Localization

In this section we introduce the partitioning approach and derive our first main results.

3.1 The Bottom Up Partitioning Approach

We say that a family $\{\mathcal{X}_1, ..., \mathcal{X}_m\}$ of nonempty disjoint subsets of $\mathcal{X}$ is a *partition* of $\mathcal{X}$, if $\mathcal{X} = \bigcup_{j=1}^m \mathcal{X}_j$. Given a probability measure ν on $\mathcal{X}$, let $p_j = \nu(\mathcal{X}_j)$. We endow each $\mathcal{X}_j$ with a probability measure by restricting the conditional probability $\nu_j(A) := \nu(A|\mathcal{X}_j) = p_j^{-1}\nu(A \cap \mathcal{X}_j)$ to the Borel sigma algebra on $\mathcal{X}_j$.

We further assume that $\mathcal{H}_j$ is a (separable) RKHS, equipped with a measurable positive semi-definite real-valued kernel K_j on each $\mathcal{X}_j$, bounded by κ_j. Note that any function in $\mathcal{H}_j$ is only defined on X_j. To make them globally defined, we extend each function $f \in \mathcal{H}_j$ to a function $\hat{f} : \mathcal{X} \longrightarrow \mathbb{R}$ by extending as the zero-function, i.e. $\hat{f}(x) = f(x)$ for any $x \in \mathcal{X}_j$ and $\hat{f}(x) = 0$ else. In particular, $\hat{K}_j$ denotes the kernel extended to $\mathcal{X}$, explicitly given by $\hat{K}_j(x, x') = K_j(x, x')$ for any $x, x' \in \mathcal{X}_j$ and zero else. Then the space $\hat{\mathcal{H}}_j := \{\hat{f} : f \in \mathcal{H}_j\}$ equipped with the norm $||\hat{f}||_{\hat{\mathcal{H}}_j} = ||f||_{\mathcal{H}_j}$ is again an RKHS of functions on $\mathcal{X}$ with kernel $\hat{K}_j$. Finally, the direct sum

$$\mathcal{H} := \bigoplus_{j=1}^m \hat{\mathcal{H}}_j = \{\, \hat{f} = \sum_{j=1}^m \hat{f}_j \;:\; \hat{f}_j \in \hat{\mathcal{H}}_j \,\}$$

with norm

$$||\hat{f}||_{\mathcal{H}}^2 = \sum_{j=1}^m p_j \, ||\hat{f}_j||_{\hat{\mathcal{H}}_j}^2$$

is also an RKHS for which

$$K(x, x') = \sum_{j=1}^m p_j^{-1} \hat{K}_j(x, x') \,, \tag{8}$$

$x, x' \in \mathcal{X}$, is the reproducing kernel, see Aronszajn (1950).

Given training data $\mathcal{D} = \{x_i, y_i\}_{i\in[n]}$, we let

$$I_j = \{i \in [n] : x_i \in \mathcal{X}_j\}$$

the set of indices indicating the samples associated to $\mathcal{X}_j$, with $|I_j| = n_j$. We split $\mathcal{D}$ according to the above partition, i.e. we let $\mathcal{D}_j = \{x_i, y_i\}_{i\in I_j}$. We further let $\mathbf{x}_j = (x_i)_{i\in I_j}$, $\mathbf{y}_j = (y_i)_{i\in I_j}$.

Fixing a regularization parameter $\lambda > 0$, we compute for each $\mathcal{D}_j$ a local KRLS estimator (compare with (2) in the global setting)

$$\hat{f}^{\lambda}_{\mathcal{D}_j} := \sum_{i\in I_j} \alpha_j^{(i)} \hat{K}_j(x_i, \cdot) \in \hat{\mathcal{H}}_j \ ,$$

where $\alpha_j \in \mathbb{R}^{n_j}$ is given by

$$\alpha_j = (\mathbb{K}_j + n_j\lambda)^{-1}\mathbf{y}_j$$

and with $\mathbb{K}_j$ the kernel matrix associated to $\mathcal{D}_j$.

Finally, the overall estimator is defined by

$$\hat{f}^{\lambda}_{\mathcal{D}} := \sum_{j=1}^{m} \hat{f}^{\lambda}_{\mathcal{D}_j} \ , \tag{9}$$

which by construction belongs $\mathcal{H}$ and decomposes according to the direct sum $\hat{\mathcal{H}}_1 \oplus ... \oplus \hat{\mathcal{H}}_m$.

3.2 Finite Sample Bounds

Our aim is to give an upper bound for the expected excess risk

$$\mathbb{E}\Big[\ \mathcal{E}(\hat{f}^{\lambda}_{\mathcal{D}}) - \mathcal{E}(f_\rho)\ \Big] \ .$$

In view of the regularity assumptions made in the global setting and described in the previous section, it is now straightforward how to express local regularity:

Assumption 2 (Regularity)**.** *1. The regression function f_ρ belongs to $\mathcal{H}$ and thus has a unique representation $f_\rho = f_1 + ... + f_m$, with $f_j \in \hat{\mathcal{H}}_j$.*

2. The local regularity of the regression function is measured in terms of a source condition:

$$||T_j^{-r_j} f_j||_{\hat{\mathcal{H}}_j} \le R \ , \qquad 0 < r_j \le \frac{1}{2} \ , \tag{10}$$

with $R < \infty$.

Note that this Assumption implies a global regularity of f_ρ as

$$||T^{-r} f||_{\mathcal{H}} \le R \ , \quad 0 < r \le \frac{1}{2} \ ,$$

with $r = \min(r_1, ..., r_m)$ and $R < \infty$.

Furthermore, we need some compatibility between the local effective dimensions and the global effective dimension in terms of the local probabilities p_j .

Assumption 3 (Capacity)**.** *1. The local effective dimensions obey*

$$m \sum_{j=1}^{m} p_j \, \mathcal{N}(T_j, \lambda) = \mathcal{O}(\mathcal{N}(T, m\lambda)) \ . \tag{11}$$

2. Global capacity: For some $0 < \gamma \le 1$

$$\mathcal{N}(T, \lambda) \lesssim \lambda^{-\gamma} \ .$$

Eq. (11) is in particular an exact equality if $p_j \equiv \frac{1}{m}$.

As in the global learning problem, the choice of the regularization parameter $\lambda = \lambda_n$ depending on the sample size n is crucial for the algorithm to work well. Interestingly, our main result shows that choosing λ locally on each subset exactly in the same way as for the global learning KRLS problem (see (7)) leads to the same error bounds as in (6).

Theorem 1 (Finite Sample Bound)**.** *Let $n_j = \lfloor \frac{n}{m} \rfloor$. Then, with the choice*

$$\lambda_n \simeq \left(\frac{1}{n}\right)^{\frac{1}{2r+1+\gamma}} \tag{12}$$

and with

$$m \lesssim n^{\alpha} \ , \quad \alpha \le \frac{2r}{2r+1+\gamma} \ . \tag{13}$$

we have the following error bound

$$\mathbb{E}\Big[\ \mathcal{E}(\hat{f}^{\lambda}_{\mathcal{D}}) - \mathcal{E}(f_\rho)\ \Big] \lesssim R^2 \left(\frac{1}{n}\right)^{\frac{2r+1}{2r+1+\gamma}} \ . \tag{14}$$

Condition (13) tells us that the sample size needs to be large enough on each local set in order to guarantee meaningful bounds. We can see that *large enough* depends here on the regularity r and the capacity γ.

We emphasize that the rhs of (14) coincides (for the case $m = 1$) with the minimax optimal rate of convergence, as shown in Caponnetto and De Vito (2006) and Blanchard and Mücke (2017) . Note that for $m > 1$ there is no explicit proof of lower bounds available in the literature (because of our additional hypothesis (11) , restricting the considered model class).

3.3 Incorporating Locality: Improved Error Bounds

Our result in Theorem 1 shows that the error bound is indeed guided by the lowest degree of regularity. Next

we show, that sometimes we can do even better if low regularity only occurs on a local set having small probability. To be more precise, assume that there is an exceptional set E of indices such that the smoothness of f_ρ is low on each set $\mathcal{X}_j$, $j \in E$ and higher on each $\mathcal{X}_j$, $j \in E^c$. For ease of reading we shall only analyze the most simple case given by:

Assumption 4 (Regularity). *There are $r_l, r_h \in (0, \frac{1}{2}]$, with $r_l < r_h$ (corresponding to low smoothness and high smoothness) and there are $R_l < \infty$, $R_h < \infty$ such that*

$$||T_j^{-r_l} f_j||_{\hat{\mathcal{H}}_j} \leq R_l \ , \quad \forall j \in E \ ,$$

$$||T_j^{-r_h} f_j||_{\hat{\mathcal{H}}_j} \leq R_h \ , \quad \forall j \in E^c \ .$$

Furthermore, assume that for any n sufficiently large

$$\left(\sum_{j \in E} p_j \right) \lesssim \left(\frac{R_h}{R_l} \right)^2 \lambda_n^{2(r_h - r_l)} \ .$$

Here, λ_n is given by (15).

Thus, global smoothness is given by the small degree r_l, while local smoothness on the complement of the exceptional set is higher. We emphasize that this is an additional assumption on the sampling distribution ν. Assumption 4 then ensures that the probability of the exceptional set is so small that the error bound will actually be governed by the higher smoothness r_h, leading to an improved finite sample bound. More precisely,

Theorem 2 (Improved error Bound). *Let $n_j = \lfloor \frac{n}{m} \rfloor$. Then, with the choice*

$$\lambda_n \simeq \left(\frac{1}{n} \right)^{\frac{1}{2r_h + 1 + \gamma}} \tag{15}$$

and with

$$m \lesssim n^\alpha \ , \quad \alpha \leq \frac{2r_h}{2r_h + 1 + \gamma} \ . \tag{16}$$

we have the following improved error bound

$$\mathbb{E}\Big[\ \mathcal{E}(\hat{f}_{\mathcal{D}}^{\lambda}) - \mathcal{E}(f_\rho) \ \Big] \lesssim \ R_h^2 \left(\frac{1}{n} \right)^{\frac{2r_h + 1}{2r_h + 1 + \gamma}} \ . \tag{17}$$

Again, for giving meaningful bounds the sample size needs to be large enough on each local set, depending on the regularity r and capacity γ.

4 KRLS Nyström Subsampling

In this section we recall the popular KRLS Nyström subsampling method. For simplicity, we restrict ourselves to so called *Plain Nyström*, which works as follows: Given a training set $x_1, ..., x_n$ of random inputs, we sample uniformly at random without replacement $l \leq n$ points $\tilde{x}_1, ..., \tilde{x}_l$. Now the crucial idea is to seek for an estimator for the unknown f_ρ in a reduced space

$$\mathcal{H}_l = \{ \ f \ : \ f = \sum_{j=1}^{l} \alpha_j K(\tilde{x}_j, \cdot) \ , \ \alpha \in \mathbb{R}^l \ \} \, .$$

In Rudi et al. (2015) it is shown that the solution of the minimization problem

$$\min_{f \in \mathcal{H}_l} \ \frac{1}{n} \sum_{j=1}^{n} (f(x_j) - y_j)^2 + \lambda ||f||_{\mathcal{H}_l}^2$$

is given by

$$\hat{f}_{n,l}^{\lambda} = \sum_{j=1}^{l} \alpha_j K(\tilde{x}_j, \cdot) \ , \tag{18}$$

with

$$\alpha = (\mathbb{K}_{nl}^* \mathbb{K}_{nl} + n\lambda \mathbb{K}_{ll})^\dagger \ \mathbb{K}_{nl}^* \mathbf{y} \ ,$$

where $(\mathbb{K}_{nl})_{ij} = K(x_i, \tilde{x}_j)$, $(\mathbb{K}_{ll})_{kj} = K(\tilde{x}_k, \tilde{x}_j)$, $i = 1, ..., n$, $k, j = 1, ..., l$ and $A^\dagger$ denotes the generalized inverse of a matrix A.

Clearly, one aims at minimizing the number l of subsamples needed for preserving minimax optimality. We amplify the results in Rudi et al. (2015) by explicitly computing how l needs to grow when the total number of samples n tends to infinity. We exhibit the explicit dependence on the regularity parameter r and on the capacity assumption, parametrized by γ. Furthermore, we refine the analysis in Rudi et al. (2015) by deriving bounds in expectation removing the dependence of l on the confidence level. This will be crucial for deriving our optimality results in the next section.

We consider the setting of Section 3 with $m = 1$. Granted Assumptions 2 and 3, one has:

Theorem 3 (KRLS-Plain Nyström). *If the number l of subsampled points satisfies*

$$l \gtrsim n^\beta \ , \quad \beta \geq \frac{1 + \gamma}{2r + 1 + \gamma} \ , \tag{19}$$

and if $r \in [0, \frac{1}{2}]$ then the choice (12) *for $(\lambda_n)_n$ leads to the error bound*

$$\mathbb{E}\Big[\mathcal{E}(\hat{f}_{n,l_n}^{\lambda_n}) - \mathcal{E}(f_\rho) \Big] \lesssim R^2 \left(\frac{1}{n} \right)^{\frac{2r+1}{2r+1+\gamma}} \ .$$

Our main result shows that the number of subsampled points can be substantially reduced from $\mathcal{O}\big(n^\beta \log(n)\big)$, see Rudi et al. (2015), to actually $\mathcal{O}\big(n^\beta\big)$. Figure 1 illustrates the interplay between subsampling level l and regularization parameter λ on an artificial dataset using a sobolev type kernel. If sufficient datapoints have been subsampled, then the plot shows a clear minimum of the mean squared error for a certain λ.

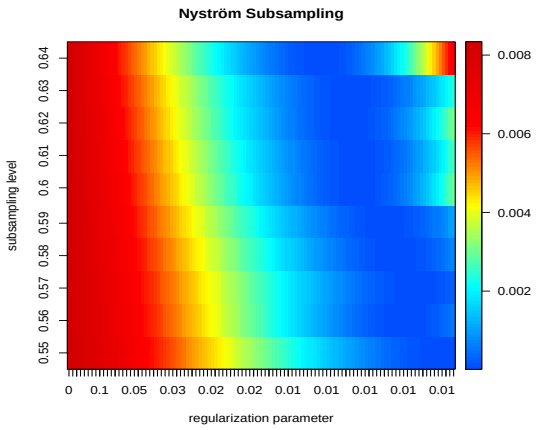

Figure 1: Interplay between subsampling level l and regularization parameter λ on an artificial dataset. The plot shows the mean squared error.

5 Combining Localization and Subsampling

In this section we establish, that upper rates of convergence are preserved if one combines the partitioning approach of Section 3 with the Nyström subsampling approach of the previous section. For simplicity we assume that the local sample size is roughly the same on each partition, i.e satisfies $n_j = \lfloor \frac{n}{m} \rfloor$ and that the number $l = l_n$ of subsample points also is equal on each subsample.

For $j = 1, ..., m$, and $1 \leq l \leq \frac{n}{m}$ let $\tilde{I}_{j,l} := \{i_{j,1}, ..., i_{j,l}\} \subseteq I_j$, with I_j as above ($\tilde{I}_{j,l}$ denotes the set of indices of subsampled inputs on each $\mathcal{X}_j$). For each subsample $\mathcal{D}_j$, with a regularization parameter $\lambda > 0$, we compute a local estimator

$$\hat{f}^{\lambda}_{\mathcal{D}_j} := \sum_{i \in \tilde{I}_{j,l}} \alpha_j^{(i)}(\lambda) \hat{K}_j(x_i, \cdot) \in \hat{\mathcal{H}}_{j,l} \, ,$$

where $\alpha_j \in \mathbb{R}^{\frac{n}{m}}$ is given in (18), with n replaced by $\frac{n}{m}$. The overall estimator is constructed as above and defined by

$$\hat{f}^{\lambda}_{\mathcal{D}} := \sum_{j=1}^{m} \hat{f}^{\lambda}_{\mathcal{D}_j} \, , \tag{20}$$

which by construction decomposes according to the direct sum $\mathcal{H} = \hat{\mathcal{H}}_1 \oplus ... \oplus \hat{\mathcal{H}}_m$. Then we have:

Theorem 4. *Let $r = \min(r_1, ..., r_m)$. If the number l of subsampled points on each local set satisfies*

$$l \sim n^{\beta} \, , \quad \beta = \frac{1+\gamma}{2r+1+\gamma} \, , \tag{21}$$

and if the number of local sets satisfies

$$m \lesssim n^{\alpha} \, , \quad \alpha \leq \frac{2r}{2r+1+\gamma} \, ,$$

then the choice (12) *for the regularization parameter λ_n guarantees the error bound*

$$\mathbb{E}\Big[\mathcal{E}(\hat{f}^{\lambda_n}_{\mathcal{D}}) - \mathcal{E}(f_\rho)\Big] \lesssim R^2 \left(\frac{1}{n}\right)^{\frac{2r+1}{2r+1+\gamma}} \, , \tag{22}$$

provided n is sufficiently large.

Clearly, as in Theorem 2, a version of the above result still holds if global smoothness is violated on an exceptional set E of small probability as amplified in Assumption 4. We leave a precise formulation (and its proof) to the reader.

6 Discussion and Comparison to other Approaches

First results establishing learning rates using a KRLS partition-based approach for smoothness parameter $r = 0$ and polynomially decaying eigenvalues are given in Tandon et al. (2016). The authors establish upper rates of convergence under an additional assumption on the probability of the local sets $\mathcal{X}_j$, requiring the existence of sufficiently high moments in $L^2(\nu)$ of the eigenfunctions of their local covariance operators, uniformly over all subsets, in the limit $n \to \infty$. However, while the decay rate of the eigenvalues can be determined by the smoothness of K (see e.g. Ferreira and Menegatto (2009) and references therein) it is a widely open question which (general) properties of the kernel imply such assumptions on the eigenfunctions. We remove these assumptions on the eigenfunctions of the covariance operator which are restrictive and difficult to prove. In addition, we allow locally different degrees of smoothness, improving finite sample bounds.

The paper Meister and Steinwart (2016) considers localized SVMs, localized tuned Gaussian kernels and a corresponding direct sum decomposition, where a global smoothness assumption is introduced in terms of a scale of Besov spaces. Instead of using the effective dimension $\mathcal{N}(\lambda)$ as a measure for complexity, the authors use entropy numbers, obtaining minimax optimal rates. We extend these results by going beyond Gaussian kernels and allowing more general input spaces than open subsets of $\mathbb{R}^d$, allowing in addition the choice of different local kernels.

We also compare the partitioning approach with distributed learning (parallelizing) for KRLS, as recently analyzed in Guo et al. (2017) and Mücke and Blanchard (2018). The distributed learning algorithm is based on a uniform partition of the given data set

$$\mathcal{D} = \{(X_1, Y_1), ..., (X_n, Y_n)\} \subset \mathcal{X} \times \mathbb{R}$$

into m disjoint equal-size subsets $\mathcal{D}_1, ..., \mathcal{D}_m$. On each subset $\mathcal{D}_j$, one computes a local estimator $\hat{f}^{\lambda}_{D_j}$ us-

Table 1: Computational Cost

KRLS	$\mathcal{O}(n^3)$
localized KRLS	$\mathcal{O}\big((\frac{n}{m})^3\big),\ 1 \leq m \leq n^\alpha$
Nyström	$\mathcal{O}(nl^2 + l^3),\ n^\beta \leq l \leq n$
local Nys.	$\mathcal{O}(\frac{n}{m}l^2 + l^3),\ n^\beta \leq l \leq \frac{n}{m}$
distributed KRLS	$\mathcal{O}\big((\frac{n}{m})^3\big),\ 1 \leq m \leq n^\alpha$

ing KRLS (or more general, a spectral regularization method). The final estimator is given by simple averaging: $\bar{f}_D^\lambda := \frac{1}{m}\sum_{j=1}^m \hat{f}_{D_j}^\lambda$.

In this setting, one takes a similiar point of view as in our main Theorem 1. Both, Guo et al. (2017) and Mücke and Blanchard (2018) provide an answer to the question: How much is the number m of local machines allowed to grow with the sample size n in order to preserve minimax optimal rates of convergence? It has been shown by these authors, that

$$m_n \sim n^\alpha\ , \qquad \alpha \leq \frac{2r}{2r+1+\gamma}$$

gives a sufficient condition. Here, $r \in (0, \frac{1}{2}]$ is again the regularity parameter of the objective function and $0 < \gamma \leq 1$ characterizes the decay of the effective dimension. Note that this relation between sample size n and number m of subsamples precisely agrees with our equation (13) . We have condensed the computational cost of all these methods in Table 1.

7 Conclusion

We have shown that the twofold effect of partitioning and subsampling may substantially reduce computational cost, if the number of local sets is sufficiently small w.r.t. the amount of data at hand and if the number of subsampled inputs is sufficiently large w.r.t. the sample size. In both cases we were able to improve or amplify the existing results. Furthermore, we derived a rigorous version of the principle *In partitioning, low smoothness on exceptional sets of small probability does not affect finite sample bounds.*

Acknowledgements

The author acknowledges support by the German Research Foundation under DFG Grant STE 1074/4-1. Furthermore, the author is grateful to Markus Klein for useful discussions.

References

N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68 (3):337–404, 1950.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. *JMLR Workshop and Conference Proceedings*, 30, 2013.

G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 2017. doi:10.1007/s10208-017-9359-7.

A. Caponnetto and E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2006.

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Analysis and Applications*, 08, No. 02, 2010.

F. Cucker and S. Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2(4):413–428, 2002.

L. H. Dicker, D. P. Foster, and D. Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electron. J. Statist.*, 11(1):1022–1047, 2017. doi: 10.1214/17-EJS1258.

J. C. Ferreira and V. A. Menegatto. Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral equations and Operator Theory*, 64, 2009.

S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv:1702.07254*, 2017.

Z.-C. Guo, S.-B. Lin, and D.-X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

J. Lin, A. Rudi, L. Rosasco, and V. Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 2018. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2018.09.009.

M. Meister and I. Steinwart. Optimal learning rates for localized svms. *Journal of Machine Learning Research*, 17(194):1–44, 2016.

N. Mücke and G. Blanchard. Parallelizing spectrally regularized kernel algorithms. *Journal of Machine Learning Research*, 19(30):1–29, 2018.

L. Rosasco, E. De Vito, and A. Verri. Spectral methods for regularization in learning theory. Technical Report 05-18, UniversitÃ di Genoa, DISI, 2005.

A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems 28*, 2015.

N. Segata and E. Blanzieri. Fast and scalable local kernel machines. *Journal of Machine Learning Research*, 11:1883–1926, 2010.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

R. Tandon, S. Si, and P. Ravikumar. Kernel ridge regression via partitioning. arXiv Preprint (1608.01976), 2016.

C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems 13*, 2000.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9):2077–2098, 2005. ISSN 0899-7667. doi: 10.1162/0899766054323008. URL `http://dx.doi.org/10.1162/0899766054323008`.