# Fast Statistical Leverage Score Approximation in Kernel Ridge Regression

**Yifan Chen**
UIUC
yifanc10@illinois.edu

**Yun Yang**
UIUC
yy84@illinois.edu

## Abstract

Nyström approximation is a fast randomized method that rapidly solves kernel ridge regression (KRR) problems through sub-sampling the $n$-by-$n$ empirical kernel matrix appearing in the objective function. However, the performance of such a sub-sampling method heavily relies on correctly estimating the statistical leverage scores for forming the sampling distribution, which can be as costly as solving the original KRR. In this work, we propose a linear time (modulo poly-log terms) algorithm to accurately approximate the statistical leverage scores in the stationary-kernel-based KRR with theoretical guarantees. Particularly, by analyzing the first-order condition of the KRR objective, we derive an analytic formula, which depends on both the input distribution and the spectral density of stationary kernels, for capturing the non-uniformity of the statistical leverage scores. Numerical experiments demonstrate that with the same prediction accuracy our method is orders of magnitude more efficient than existing methods in selecting the representative sub-samples in the Nyström approximation.

## 1 Introduction

The major computational bottleneck of kernel-based machine learning methods, such as kernel ridge regression (KRR) (Shawe-Taylor et al., 2004; Hastie et al., 2005), lies in the calculation of certain matrix inverse involving an $n$-by-$n$ symmetric and positive semidefinite (PSD) empirical kernel matrix $K_n \in \mathbb{R}^{n\times n}$ over $n$ inputs in the $d$-dimensional Euclidean space $\mathbb{R}^d$. For most common kernels, the empirical kernel matrix $K_n$ is nearly singular with its effective rank being captured by the so-called *effective statistical dimension* (Alaoui and Mahoney, 2015; Yang et al., 2017b) $d_{stat}$ that is problem-dependent and can be substantially smaller than the sample size $n$. For example, under the optimal choice of the regularization parameter, if the kernel function is a Matérn kernel with smoothness parameter $\nu > 0$, then the statistical dimension in the KRR is $d_{stat} = \mathcal{O}(n^{\frac{d}{2\nu+2d}})$ (Kanagawa et al., 2018).

### 1.1 Related Works

Due to this intrinsic low-rankness of $K_n$, several existing papers developed randomized algorithms, such as the Nyström method (Alaoui and Mahoney, 2015), randomized sketches (Yang et al., 2017b; Ahle et al., 2020), random Fourier features (Rahimi and Recht, 2008; Avron et al., 2017), and its improvement quadrature Fourier features (Mutny and Krause, 2018), for obtaining a rank $\widetilde{\mathcal{O}}(d_{stat})$ ($\widetilde{\mathcal{O}}(\cdot)$ means $\mathcal{O}(\cdot)$ modulo poly-log terms) approximation of $K_n$. In particular, sampling-based algorithms, such as the Nyström method, avoid explicitly constructing the $n$-by-$n$ matrix $K_n$, and only require $\widetilde{\mathcal{O}}(nd_{stat})$ evaluations of the kernel function. This property is particularly appealing since both the time and space complexity can be reduced to even below the $\mathcal{O}(n^2)$ benchmark complexity of constructing and storing the empirical kernel matrix. From an algorithmic perspective, *statistical leverage scores*, as a measure of the structural non-uniformity of the inputs in forming $K_n$, can be used for constructing an importance sampling distribution that leads to high-quality low-rank approximations in the Nyström method. We refer the readers to some recent papers (Mahoney et al., 2011; Drineas et al., 2012) for more details.

However, the exact computation of the statistical leverage scores bears the same $\mathcal{O}(n^3)$ time complex-

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

ity and $\mathcal{O}(n^2)$ space complexity (Mahoney et al., 2011) as inverting the $n$-by-$n$ empirical kernel matrix. Researchers thus turn to the question of whether there is an efficient and accurate method of approximately computing the leverage scores. For example, some works (Alaoui and Mahoney, 2015; Rudi et al., 2015) borrowed the random projection idea (Drineas et al., 2012) in designing an approximation algorithm for computing the statistical leverage scores in the Nyström method in the context of KRR. Their algorithm has a worst case time complexity $\mathcal{O}\big(\frac{n^3}{d_{stat}^2}\big)$ that may exceed the $\mathcal{O}(nd_{stat}^2)$ complexity in subsequent steps for small $d_{stat}$. As a refinement, Musco and Musco (2017) developed Recursive-RLS, a recursive version of the prior algorithm (Alaoui and Mahoney, 2015) with overall time complexity $O(nd_{stat}^2)$ by alternating between updating the statistical leverage scores and drawing new subsamples based on the current scores. SQUEAK (Calandriello et al., 2017) adapts the algorithm to an online setting, attaining the same accuracy and having the same complexity order with only one pass over the data. BLESS (Rudi et al., 2018) adopts a path-following algorithm that further reduces the subsampling time complexity to $\mathcal{O}(\min(\frac{1}{\lambda}, n)d_{stat}^2 \log^2 \frac{1}{\lambda})$ where $\lambda$ is the regularization parameter in the KRR. With the choice of $\lambda = \mathcal{O}(\frac{d_{stat}}{n})$ that leads to the optimal error rate, the complexity of BLESS would be $\widetilde{\mathcal{O}}(nd_{stat})$.

### 1.2 Our Contribution

Most previous algorithms are algebraic methods by approximating matrix operations and apply to any positive semidefinite (PSD) kernel. In this work, we focus on stationary kernels and follow a completely different route of utilizing large sample properties of KRR to develop a new analytical approach for approximating the statistical leverage scores. Under a classical nonparametric setting, the new approach requires $\widetilde{\mathcal{O}}(n)$ time and space complexity, and provably also attains the optimal statistical accuracy in the KRR. In a nutshell, rather than approximating the leverage scores by preconstructing a low-rank approximation to $K_n$ (Drineas et al., 2012), our method uses structural information contained in the kernel function and the underlying input distribution to infer how other inputs influence the statistical leverage score at a given location. In particular, we derive an explicit and computable formula,

$$\int_{\mathbb{R}^d} \frac{1}{p(x_i) + \lambda/m(s)} \, \mathrm{d}s, \quad \forall i \in [n],$$

where $m(\cdot)$ is the spectral density function of the stationary kernel we use, and $p(x_i)$ is the density of the input $x_i$. This formula is applied to approximate the rescaled statistical leverage score, which is proportional to the true statistical leverage score, at each observed point. We also provide the theoretical guarantees that the approximation formula has a vanishing relative error as $n \to \infty$.

Our development is based on the existing works (Silverman, 1984; Yang et al., 2017a) on the equivalent kernel representation of the KRR solution. The consequent theory sheds some light on the behaviour of leverage scores, and a simple application is the following rule of thumb: for the Matérn kernel with smoothness $\nu$, the statistical leverage score at point $x$ in $\mathbb{R}^d$ is proportional to $\min\{1, (\lambda/p(x))^{1-d/(2\nu+d)}\}$, where the regularization parameter $\lambda = \Theta(d_{stat}/n)$. This scaling indeed matches the previous research on the asymptotic equivalent of the regularized Christoffel function (Pauwels et al., 2018), which has intrinsic connections with statistical leverage scores. We also show through numerical experiments that our method exhibits encouraging performance compared to other methods, which further improves the overall runtime of KRR.

## 2 Background and Problem Formulation

In this section, we set up the notation and briefly introduce the background. We begin with a short review on reproducing kernel Hilbert space (RKHS) and kernel ridge regression (KRR). After that, we invoke a useful and important formula that represents the norm of any RKHS induced by a stationary kernel via Fourier transforms and Parseval's identity. Then we describe the class of Nyström -method-based approaches as our primary focus for approximately computing the KRR. Finally, we introduce the notation of equivalent kernel that plays important roles in both understanding the theoretical properties of the KRR and motivating our proposed method.

### 2.1 Reproducing kernel Hilbert space and kernel ridge regression

**Reproducing kernel Hilbert space.** Particularly, any RKHS is generated by a PSD kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and there exists a correspondence between any RKHS (or its induced norm $\|\cdot\|_{\mathbb{H}}$) and its reproducing kernel (see the books (Berlinet and Thomas-Agnan, 2011; Wahba, 1990; Gu, 2013) for more details). Most widely used kernels are stationary, meaning that $K(x, y)$ depends on $x$ and $y$ only through their difference $(x-y)$. Due to this definition, we may abuse the notation $K(u)$ to mean $K(x, x+u)$ for any $x$. We will make this stationary kernel assumption throughout the paper. More specifically, we primarily focus on Matérn kernels, although the development can be

straightforwardly extended to other stationary ones.

**Kernel ridge regression.** Consider a dataset $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ consisting of $n$ pairs of points in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input (predictor) space and $\mathcal{Y} \subseteq \mathbb{R}$ the response space. To characterize the dependence of the response on the predictor, we assume the following standard nonparametric regression model as the underlying data generating model, $y_i = f^*(x_i) + \varepsilon_i$, $i = 1, 2, \ldots, n$, where $f^* : \mathcal{X} \to \mathcal{Y}$ is the unknown regression function to be estimated, and the random noises $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, or any sub-Gaussian distribution with mean zero and variance $\sigma^2$. Under the common regularity assumption that the true regression function $f^*$ belongs to an RKHS $\mathbb{H}$, it is natural to estimate $f^*$ by an estimator $\widehat{f}$, which minimizes the sum of a least-squares goodness-of-fit term and a penalty term involving the squared norm $\|\cdot\|_{\mathbb{H}}$ associated with $\mathbb{H}$. This leads to the following estimating procedure known as *kernel ridge regression* (KRR) (Shawe-Taylor et al., 2004; Hastie et al., 2005),

$$\widehat{f} = \arg\min_{f \in \mathbb{H}} \Big\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathbb{H}}^2 \Big\}. \tag{1}$$

The optimization problem (1) appears to be infinite-dimensional over a function space $\mathbb{H}$, while indeed its solution can be obtained by solving an $n$-dimensional quadratic program thanks to the representer theorem (Kimeldorf and Wahba, 1971). More precisely, for any two $\mathcal{X}$-valued vectors $a = (a_1, \ldots, a_p)^T \in \mathcal{X}^p$ and $b = (b_1, \ldots, b_q)^T \in \mathcal{X}^q$, we use the notation $K(a, b)$ to denote the $p$-by-$q$ matrix whose $(i, j)$-th component is $K(a_i, b_j)$ for $i \in [p]$ and $j \in [q]$. Let $X_n = (x_1, \ldots, x_n)^T \in \mathcal{X}^n$ and $Y_n = (y_1, \ldots, y_n)^T \in \mathcal{Y}^n$. Under this notation, the solution $\widehat{f}$ takes the form as

$$\begin{aligned} \widehat{f}(x) &:= K(x, X_n) \left(K_n + n\lambda I_n\right)^{-1} Y_n \\ &= \frac{1}{n} \sum_{i=1}^n G_\lambda(x, x_i)\, y_i, \end{aligned} \tag{2}$$

where $K_n := K(X_n, X_n)$ is the $n$-by-$n$ empirical kernel matrix. Here, the weight function $G_\lambda : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ characterizes the impact of each observed pair $(x_i, y_i)$ on $\widehat{f}(x)$, and plays an important role in determining the optimal importance sampling weights in the Nyström method described. One key observation is that the weight function $G_\lambda$ depends on the dataset $\mathcal{D}_n$ only through the design points $\{x_i\}_{i=1}^n$ and the regularization parameter $\lambda$ (which usually depends on the sample size $n$). This fact leads to the development of equivalent kernel approximation, as we will come back shortly in Section 2.4.

A final remark of the subsection is that solving for $\widehat{\omega}$ requires time complexity $\mathcal{O}(n^3)$ of inverting an $n$-by-$n$ matrix, which becomes formidable when $n$ gets large. The practical demand of computationally scalable methods for implementing the KRR results in a large volume of literature on approximation algorithms, including our current work.

### 2.2 The relation between RKHS norm and Fourier transform

In this subsection, we describe a useful representation theorem that characterizes the RKHS of a stationary kernel function via the Fourier transform. Before formally introducing this theorem, we set up some notation. We use $L_p(\mathbb{R}^d) = \big\{f : \mathbb{R}^d \to \mathbb{R}, \int_{\mathbb{R}^d} \big|f(x)\big|^p \,\mathrm{d}x < \infty\big\}$ to denote the space of all $L_p$ integrable functions over $\mathbb{R}^d$ for $p \geq 1$. For any function $f \in L_1(\mathbb{R}^d)$, we use $\mathscr{F}[f]$ to denote its Fourier transform defined by $\mathscr{F}[f](s) = \int_{\mathbb{R}^d} f(x)\, e^{-2\pi\sqrt{-1}x^T s} \,\mathrm{d}x$, for all $s \in \mathbb{R}^d$, and $\mathscr{F}^{-1}[g]$ the inverse transform of a function $g$ in the frequency domain as $\mathscr{F}^{-1}[g](x) = \int_{\mathbb{R}^d} g(s)\, e^{2\pi\sqrt{-1}x^T s} \,\mathrm{d}s$, for all $x \in \mathbb{R}^d$. We use $\bar{z}$ to denote the complex conjugate of any $z \in \mathbb{C}$, the space of complex numbers. The classical Bochner's theorem shows that the Fourier transform of any PSD and stationary kernel function is nonnegative.

The following theorem, which is not new but less well-known in the machine learning literature, provides a characterization of the RKHS with kernel $K$ through its spectral density function. Similar statements can be found in two previous papers (Wendland 2004, Thm 10.12 and Belkin 2018, Appendix A). For the sake of completeness, we also include a proof in Section 6.2 in the appendix, which is motivated by Fukumizu (2008).

**Theorem 1** (Fourier representation of RKHS)**.** *Let function $m$ be the spectral density of a PSD and stationary kernel $K$, and $\mathbb{H}$ the associated RKHS. For any $f, g \in \mathbb{H}$, we have*

$$\begin{aligned} \|f\|_{\mathbb{H}}^2 &= \int_{\mathbb{R}^d} \frac{\big|\mathscr{F}[f](s)\big|^2}{m(s)} \,\mathrm{d}s, \\ \text{and} \quad \langle f, g\rangle_{\mathbb{H}} &= \int_{\mathbb{R}^d} \frac{\mathscr{F}[f](s) \cdot \overline{\mathscr{F}[g](s)}}{m(s)} \,\mathrm{d}s. \end{aligned} \tag{3}$$

*In particular, the RKHS can be represented by $\mathbb{H} = \big\{f : \|f\|_{\mathbb{H}}^2 = \int_{\mathbb{R}^d} \big|\mathscr{F}[f](s)\big|^2 / m(s) \,\mathrm{d}s < \infty\big\}$.*

### 2.3 Nyström methods with importance subsampling

From expression (2) of the KRR solution $\widehat{f}$, the $\mathcal{O}(n^3)$ computational bottleneck comes from the inversion of the $n$-by-$n$ matrix $(K_n + n\lambda I)$. When design points $\{x_i\}_{i=1}^n$ are distinct and sample size $n$ is large, the

empirical kernel matrix $K_n$ often has a high condition number and is nearly low-rank. In particular, several recent studies (Alaoui and Mahoney, 2015; Yang et al., 2017b) show that both the computational and the statistical hardness of the KRR are captured by a quantity called the *statistical dimension* defined as

$$d_{stat} := \mathrm{Tr}\big(K_n(K_n + n\lambda I_n)^{-1}\big) = \frac{1}{n}\sum_{i=1}^{n} G_\lambda(x_i, x_i), \tag{4}$$

where $\mathrm{Tr}(A)$ means the trace of a matrix $A$.

The statistical dimension $d_{stat}$ approximately counts the number of eigenvalues of the rescaled kernel matrix $n^{-1}K_n$ whose values are above the threshold $\lambda$. Therefore, computing $\widehat{f}$ roughly amounts to solving an $d_{stat}$-dim quadratic program, and for most stationary kernels $d_{stat}$ would be much smaller than $n$. For example, for a Matérn kernel with smoothness parameter $\nu$ being a positive half integer, $d_{stat} = \mathcal{O}(n^{\frac{d}{2\nu+2d}})$ (Tuo et al., 2020). Due to the intrinsic low-rankness of $K_n$, the so-called Nyström method, which replaces $K_n$ with its low-rank approximation $L_n$, has been used for approximately solving the KRR (Gittens and Mahoney, 2016; Kumar et al., 2009; Williams and Seeger, 2001). Specially, the Nyström approximation of $K_n$ is the matrix $L_n = K_n S(S^T K_n S)^\dagger S^T K_n$, where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix $A$, and $S \in \mathbb{R}^{n\times d_{sub}}$ is a zero-one subsampling matrix whose columns are a subset of the columns in $I_n$, indicating which $d_{sub}$ observations have been selected. We use $\widehat{f}_{L_n}$ to denote the approximate KRR solution obtained by replacing $K_n$ with $L_n$ in expression (2).

Following the paper (Alaoui and Mahoney, 2015), we will refer to the diagonal elements of matrix $K_n(K_n + n\lambda I_n)^{-1}$ as the *statistical leverage scores* $\{\ell_i\}_{i=1}^n$ associated with the kernel matrix $K_n$. It is straightforward to verify from identity (2) that the $i$-th diagonal element of $K_n(K_n+n\lambda I_n)^{-1}$ is precisely $\frac{1}{n}G_\lambda(x_i, x_i)$, and thus we call $G_\lambda(x_i, x_i)$ the rescaled statistical leverage score.

The next result (Alaoui and Mahoney, 2015, Theorem 3) (after adapting to our notation) shows that if we use a randomized construction of $S$ by sampling $d_{sub} = \mathcal{O}(d_{stat}\log(n))$ columns from $I_n$ with a proper distribution $\{q_i\}_{i=1}^n$ over $[n]$ approximately proportional to the statistical leverage scores, the resulting approximate solution $\widehat{f}_{L_n}$ would attain the same statistical in-sample risk (up to a constant) as the original KRR solution $\widehat{f}$. Here, we define the in-sample prediction risk for any regression function $f$ as $R_n(f) = \|f - f^*\|_n^2 := n^{-1}\sum_{i=1}^n \big(f(x_i) - f^*(x_i)\big)^2$.

**Theorem 2** (Nyström approximation accuracy)**.** *Fix $\rho \in (0, 1/2)$. Let $L_n$ be the Nyström approximation of $K_n$ with $S$ being formed by choosing $d$ columns randomly with replacement from the columns of the identity matrix $I_n$ according to a probability distribution $\{q_i\}_{i=1}^n$. Suppose there exists some $\beta \in (0, 1]$ such that $q_i \geq \beta\, G_\lambda(x_i, x_i)/\sum_{i=1}^n G_\lambda(x_i, x_i)$, and*

$$d_{sub} \geq C_1 \frac{d_{stat}}{\beta}\log\Big(\frac{n}{\rho}\Big) \quad \text{and} \quad \lambda \geq \frac{C_2}{\min_i G_\lambda(x_i, x_i)},$$

*where $d_{stat}$ is defined in (4). Then it holds with probability at least $1 - 2\rho$ that $R_n(\widehat{f}_{L_n}) \leq C_3 R_n(\widehat{f})$. Here $C_i$, $i = 1, 2, 3$, are absolute constants.*

This theorem indicates that the problem of approximately solving the KRR reduces to that of approximately estimating the statistical leverage scores $\{\ell_i\}_{i=1}^n$. Directly computing these leverage scores using SVD requires inverting an $n$-by-$n$ matrix and is as costly as solving the original KRR optimization (2). Finding purely numerical methods for approximating these leverage scores can also be quite challenging. For example, the approximation algorithm used in the paper (Alaoui and Mahoney, 2015) has $\mathcal{O}\Big(\frac{n^3}{d_{stat}^2}\Big)$ time complexity, which significantly exceeds the $\mathcal{O}(nd_{stat}^2)$ complexity for forming $L_n$ and solving for $\widehat{f}_{L_n}$ in the Nyström approximation when $d_{stat} \ll \sqrt{n}$ (which holds for any Mátern kernel).

### 2.4 Equivalent Kernel

Our method is initially motivated by a notion, *equivalent kernel*, which was first introduced by Silverman (1984). The author showed that in the context of smoothing spline regression, as sample size $n$ goes to infinity the weight function $G_\lambda(\cdot, \cdot)$ after a proper rescaling approaches a limiting kernel function, called the equivalent kernel. A recent work (Yang et al., 2017a, Theorem 2.1) extends the context from smoothing spline regression to general kernel ridge regression. They proved that for a general kernel $K$, under the stochastic assumption that design points $\{x_i\}_{i=1}^n$ are i.i.d. distributed according to a common distribution over $\mathcal{X}$, there exists some equivalent kernel $\bar{K}_\lambda : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, such that the KRR estimator is asymptotically the same as a simple kernel type estimator with kernel function $\bar{K}_\lambda$, that is, under a suitable choice of diminishing regularization parameter $\lambda$, the following approximation error bound holds with probability tending to one as $n \to \infty$,

$$\sup_{x\in\mathcal{X}}\Big|\widehat{f}(x) - \frac{1}{n}\sum_{i=1}^n \bar{K}_\lambda(x, x_i)\, y_i\Big| \leq \gamma_n\sqrt{\lambda}, \tag{5}$$

where $\gamma_n \to 0$ as $n \to \infty$ and $\sqrt{\lambda}$ matches the magnitude of the estimation error $\sup_{x\in\mathcal{X}}|\widehat{f}(x) - f^*(x)|$.

Formula (5) predicts a theoretical limit of the statistical leverage score $\ell_i$ as $\bar{K}_\lambda(x_i, x_i)$ that is *independent* of the design points other than $x_i$. In other words, the KRR estimator can be expressed as a linear combination of $y_i$'s, whose coefficients only depend on the corresponding design point $x_i$. Besides, under mild conditions (Yang et al., 2017a) on the design distribution, these coefficients admit a limit in probability (that can be characterized via an "equivalent kernel" function) as $n \to \infty$. The theoretical result motivates us to seek a computationally efficient method for the leverage scores through approximating these $\bar{K}_\lambda(x_i, x_i)$'s.

## 3 Leverage Score Approximation via Spectral Analysis

We now turn to the main results of this work. At a high level, we propose our method and give a sketch of the derivation via Fourier transform. We also provide the analysis of the corresponding time complexity and consider some stationary kernels to which our method would be applied. Eventually, we prove that, by taking Matérn kernels as an example, our method would attain an optimal prediction risk in the KRR.

### 3.1 The proposed algorithm and a heuristic derivation

Under the notation above, we formally propose the explicit formula of our approximation method as

$$\widetilde{K}_\lambda(x_i, x_i) = \int_{\mathbb{R}^d} \frac{1}{p(x_i) + \lambda/m(s)}\,\mathrm{d}s, \tag{6}$$

where $p(x_i)$ is the density of $x_i$ defined in Section 1.2. With Eqn (6), we would use Algorithm 1 below to approximate $G_\lambda(x_i, x_i)$ for some fixed point $x_i \in \mathbb{R}^d$.

**Algorithm 1:** Estimation of the leverage scores

**Input:** the input samples $X_n$ and the spectral density $m(\cdot)$ of the stationary kernel used

**Output:** A descrete sampling distribution $\{q_i\}_{i=1}^n$

Initialize the sampling distribution $q_i = 0, \forall i = 1, \dots, n$;

Estimate the density $p_i$ of the samples $x_i, \forall i = 1, \dots, n$;

**for** *i=1:m* **do**

  Compute the integration (6) with $p_i$, and assign the value to $q_i$

**end**

Denote $Q = \sum_{i=1}^n q_i$;

Update $q_i$ as $q_i/Q, \forall i = 1, \dots, n$;

To derive the formula (6), by setting $y_i = n$, $y_j = 0$ (for any $j \neq i$), we transform the objective value in the KRR optimization problem (1) to the following functional:

$$\begin{aligned} A_{n,x_i}(f) &= \frac{1}{2n}\sum_{j=1}^n f(x_j)^2 + \frac{1}{2}\lambda\|f\|_{\mathcal{H}}^2 - f(x_i) \\ &= \frac{1}{2}\int_{\mathbb{R}^d} f(x)^2\,\mathrm{d}F_n(x) + \frac{1}{2}\lambda\|f\|_{\mathcal{H}}^2 - f(x_i), \end{aligned}$$

for any function $f \in \mathbb{H}$, where $F_n$ denotes the empirical distribution of $\{x_i\}_{i=1}^n$ and the integral is the Riemann–Stieltjes integral. The minimizer of $A_{n,x_i}(f)$ would simply be $\widehat{f}(\cdot) = \frac{1}{n}\sum_{j=1}^n G_\lambda(\cdot, x_j)y_j = G_\lambda(\cdot, x_i)$, due to the independence of $G_\lambda$ from $\{y_i\}_{i=1}^n$. Therefore, it suffices to analyze and understand this functional $A_{n,x_i}$ for any $x_i \in \mathcal{X}$. Now we assume that there exists a nice cdf $F$ over $\mathcal{X}$ that admits a Lipschitz continuous density function denoted by $p$, so that the sup-norm $\tau(n) = \|F_n - F\|_\infty := \sup_{x\in\mathcal{X}}|F_n(x) - F(x)|$ is small. We further remark here that in the most common case, $\{x_i\}_{i=1}^n$ are i.i.d. from pdf $p$, and then $\tau(n) \le C\sqrt{\log n/n}$ holds with high probability due to the Glivenko-Cantelli theorem (Van Der Vaart and Wellner, 1996). Since $A_{n,x_i}$ is convex, finding its optimum amounts to finding the unique root of its functional derivative (such as the Gateaux derivative), which is a linear operator $DA_{n,x_i}(f) : \mathbb{H} \to \mathbb{H}$ defined at each $f \in \mathbb{H}$ as

$$\begin{aligned} DA_{n,x_i}(f)(u) = &\int_{\mathbb{R}^d} f(x)\,u(x)\,\mathrm{d}F_n(x) \\ &+ \lambda\langle f, u\rangle_{\mathbb{H}} - u(x_i), \quad \text{for any } u \in \mathbb{H}. \end{aligned}$$

Since $F_n$ can be well-approximated by $F$ under the assumption $\tau(n) \to 0$ and the solution $G_\lambda(\cdot, x_i)$ is expected to approach to a Dirac delta function centered at $x_i$ as $n \to \infty$ (which will be formalized in Section 8 in the appendix), the above derivative can thus be approximated by a simpler population-level functional

$$\begin{aligned} DA_{x_i}(f)(u) = &p(x_i)\int_{\mathbb{R}^d} f(x)\,u(x)\,\mathrm{d}x \\ &+ \lambda\langle f, u\rangle_{\mathbb{H}} - u(x_i), \quad \text{for any } u \in \mathbb{H}, \end{aligned}$$

where we replace the differential $\mathrm{d}F_n(x)$ with its local approximate $p(x_i)dx$. This new operator admits a simpler form in the frequency domain thanks to Parseval's theorem (cf. Theorem 5 and Lemma 6 in the appendix),

$$\begin{aligned} DA_{x_i}(f)(u) = \int_{-\infty}^{\infty} \Big(&p(x_i)\,\mathscr{F}[f](s) + \frac{\lambda}{m(s)}\,\mathscr{F}[f](s) \\ &- \exp\big\{-2\pi\sqrt{-1}x_i s\big\}\Big)\,\overline{\mathscr{F}[u](s)}\,\mathrm{d}s. \end{aligned}$$

Therefore, the unique root of $DA_{x_i}(\cdot)$, denoted by $\widetilde{K}_\lambda(\cdot, x_i)$, can be obtained by equating the function

inside the big parenthesis in the preceding display to zero, which is the inverse Fourier transform of

$$\frac{\exp\{-2\pi\sqrt{-1}\langle x_i, s\rangle\}}{p(x_i)+\lambda/m(s)}, \quad s\in\mathbb{R}^d, \tag{7}$$

or $\widetilde{K}_\lambda(\cdot, x_i) = \mathscr{F}^{-1}\big[\big(p(x_i)+\lambda/m(s)\big)^{-1}\big](\cdot - x_i)$ due to the translation property of the Fourier transform. Replacing $x_i$ with the $i$-th design point $x_i$ leads to the following quantity

$$\widetilde{K}_\lambda(x_i, x_i) = \int_{\mathbb{R}^d} \frac{1}{p(x_i)+\lambda/m(s)}\,\mathrm{d}s,$$

due to the inverse Fourier transform formula. We show its applications to Matérn kernels as follows.

**Example (Matérn kernels)**: Matérn family (Matérn, 2013) is a class of isotropic kernels widely used in spatial statistics. The kernel function is expressed as $C_\nu(x,y) = C_\nu(x-y) = \frac{2^{1-\nu}}{\Gamma(\nu)}(a\|x-y\|)^\nu B_\nu(a\|x-y\|)$, where $B_\nu$ is a modified Bessel function of the second kind, $\nu$ is a smoothness parameter (usually half integers), and $a>0$ a scale parameter. Here we slightly abused the notation since $C_\nu$ is stationary. An important fact about the Matérn kernel $C_\nu$ is that its associated RKHS is the $(\nu+d/2)$-th order Sobolev space (we can verify it by plugging the following Fourier transform $m_\alpha(\cdot)$ into Theorem 1). The notation $\alpha = \nu + d/2$ is hence used to denote the underlying smoothness level associated with kernel $K_\alpha := C_{\alpha-d/2}$, and the rescaled leverage approximation $\widetilde{K}_\lambda$ associated with $K_\alpha$ satisfies $\widetilde{K}_\lambda(\cdot, x_i) = \mathscr{F}^{-1}\Big[\Big(p(x_i)+\lambda D_\alpha^{-1}\big(a^2+\|s\|^2\big)^\alpha\Big)^{-1}\Big](\cdot - x_i)$, where $a=\sqrt{2\nu}, D_\alpha = \Gamma(\alpha)a^{2\alpha-1}\pi^{-d/2}/\Gamma(2\alpha-1)$. For general density function $p$, the integral formula (6) with $m=m_\alpha$ provides a rule of thumb on how the statistical leverage score depends on the local input density as $\ell_i \propto \min\{1, (\lambda/p(x_i))^{1-d/(2\alpha)}\}$, which implies a relatively large value over those under-sampled regions with small $p(x_i)$.

### 3.2 Computational complexity

To give the complexity analysis, we first stress that in our theoretical development the dimension $d$ is either fixed or at most slowly (e.g. logarithmically) increases with the sample size $n$. Beyond this setting, at least theoretically, the smallest subsampling size (via statistical dimension $d_{stat}$, which is $\mathcal{O}((\log n)^{\frac{d}{2}})$ even under a Gaussian kernel) becomes comparable to $n$, making subsampling meaningless due to the curse of dimensionality. In addition, classical nonparametric literature (Silverman, 1984; Yang et al., 2017a; van der Vaart et al., 2009) suggests that a dimension $d$ of order $o(\log n)$ is necessary to make any estimator consistent.

With the requirement on $d$ above, we claim $\widetilde{K}_\lambda(x_i, x_i)$ can be efficiently computed in $\widetilde{\mathcal{O}}(n)$ time. Specifically, the overall complexity includes two parts, numerical integration, and density estimation. A key observation here is that for both parts the error rates are only required to be sub-optimal, and $o(1)$ relative error suffices to guarantee the optimality of the error rate in the KRR. With such a high tolerance of error, the two parts above could both be implemented in $\widetilde{\mathcal{O}}(n)$ time as claimed (see Section 9, 10 in the appendix for more details).

In particular, the overall complexity can be made at most polynomial in the dimension $d$. For the integration part, we can avoid the exponential dependence on $d$ by applying a polar coordinate transformation to reduce the multivariate integral (6) to a univariate integral (c.f. Section 9 in the appendix). For density estimation, some advanced methods are able to generate $n$ density estimates at sample design points in $\mathcal{O}(nd\log n)$ time with relative approxiamtion error (difference between accurate KDE and approximation methods) of magnitude $\mathcal{O}((\log n)^{-1/2})$ (such as ASKIT (March et al., 2015, Eqn (3.3)), HBE (Charikar and Siminelakis, 2017, Theorem 12), and modified HBE (Backurs et al., 2019, Theorem 1)). (Those methods only aim to approximate the original KDE, and hence have no requirements on the density but the kernel used in KDE. The exact set of assumptions for modified HBE are provided in Section 10.1 in the appendix.) In practice, when the KRR problem of interest is not high dimensional($d=\omega(1)$), we are even able to efficiently estimate the density with the optimal error rate in $\mathcal{O}(n(\log n)^d)$ time, by some classical approaches (such as KD-tree methods (Ivezic et al., 2014), fast multipole methods (Greengard and Rokhlin, 1997), and fast Gauss transforms (Greengard and Strain, 1991)), which are empirically be even faster than the advanced KDE methods above.

As a closing of this subsection, we leave a comment regarding Gaussian kernels. It seems Gaussian kernels have a low statistical dimension $d_{stat} = \mathcal{O}((\log n)^{d/2})$, which may allow previous leverage approximation methods, such as BLESS, to have a time complexity comparable to our method. We point out here the complete expression for the scale of $d_{stat}$ should be $\mathcal{O}(\sigma^{-d}(\log(n\sigma^{2d}))^{d/2})$ (Yang et al., 2017b), where $\sigma$ is the bandwidth of the Gaussian kernel used. It implies the statistical dimension of Gaussian Kernels would actually be heavily impacted by the bandwidth $\sigma$. However, as we hope to attain the optimal error rate in KRR, we need to decrease the bandwidth $\sigma$ of Gaussian kernels to $\mathcal{O}(n^{-c})$ $(c\in(0,\frac{1}{2d}))$ to significantly enrich the associated RKHS (van der Vaart et al., 2009). As a trade-off, the magnitude of $d_{stat}$

would simultaneously be increased to a polynomial of $n$, which is comparable to the scale of $d_{stat}$ using a proper Matérn kernel. Therefore, generally Gaussian kernels cannot enable the previous leverage approximation methods to enjoy the $\widetilde{\mathcal{O}}(n)$ complexity.

### 3.3 Theoretical results

We focus on the Matérn kernel $K_\alpha$ (proof for other stationary kernels can be developed similarly) and define the effective bandwidth parameter, an important auxiliary parameter analogous to the bandwidth of a Gaussian kernel, as $h := \lambda^{1/(2\alpha)}$, which indicates the smoothness of the functions in the corresponding RKHS. When $\lambda$ is set to obtain the minimax-optimal KRR estimator $\widehat{f}$, the scale of $h$ would therefore be $\Theta(n^{-1/(2\alpha+d)})$. Our first result provides an explicit error bound on $|G_\lambda(\cdot, x_i) - \widetilde{K}_\lambda(\cdot, x_i)|$ in a neighborhood of $x_i$. Particularly, Lemma 10 in the appendix shows that the equivalent kernel $\widetilde{K}_\lambda(\cdot, x_i)$ resembles a Dirac delta function centered at $x_i$ with radius $\mathcal{O}(h)$, so it suffices to characterize its difference with $G_\lambda(\cdot, x_i)$ in the local neighborhood. The regularity assumptions of our results are listed as follows:

**Assumption 1.** *There exists a distribution $F$ whose density function $p$ is Lipschitz continuous, so that $\tau(n) = \|F_n - F\|_\infty \le C_0 h^2$ for some constant $C_0 > 0$.*

**Assumption 2.** *The density function $p$ is strictly positive at the points $x_i, i = 1, \dots, n$. Besides, for each $i$, there exists some $\delta(x_i)$, such that $p(x) \ge \frac{p(x_i)}{2}$ for all $x \in B(x_i, \delta(x_i)) := \{x; \|x - x_i\| < \delta(x_i)\}$, and $\delta(x_i) \ge Ch\log(\frac{1}{h}), \forall i \in [n]$ for some sufficiently large constant $C$ independent of $(x_i, h, n)$.*

Assumption 1 needs the empirical distribution $F_n$ to be well-approximated by a smooth cdf $F$ (c.f. Section 3.1). Assumption 2 requires $x_i$ to be a proper interior point of the support of $p$, or at least $\Theta(h\log(\frac{1}{h}))$ far away from the zero density region. These two assumptions are mild. Assumption 1 is indeed a generalized version of a common assumption "$x_i$'s are drawn iid from a distribution F with Lipschitz density p", while also compatible with fixed design setting. Assumption 1 would automatically be satisfied when the quoted assumption holds, with $\tau(n) = \mathcal{O}((d/n)^{1/2})$ guaranteed by multivariate Glivenko-Cantellli theorem (p. 828 Shorack and Wellner, 2009, Theorem 1). Another remark about Assumption 1 is that the Lipschitz continuity of the density $p$ is enough for KDE to produce consistent estimation with mean squared error $\mathcal{O}(n^{-\frac{2}{d+2}})$ (Walter et al., 1979), which is even dominated by the approximation error of the KDE approximation methods mentioned above. (Thus now we can conclude the KDE methods above with $\mathcal{O}(nd(\log n))$ time complexity could provide sufficient accuracy.) For Assumption 2, as long as $p(x_i) > 0$ and $p$ is continuous, it always holds in the large scale setting as $n \to \infty$ since $h\log(\frac{1}{h}) \to 0$ as $h \to 0$. Moreover, we only need to verify it for the observed design points $\{x_i\}_{i=1}^n$ in conjunction with Theorem 2, and by definition $p(\cdot)$ is automatically positive at these observed points.

**Theorem 3** (Leverage score approximation)**.** *If Assumptions 1 and 2 hold, then for any $i \in [n]$*

$$\sup_{x \in B(x_i, \delta(x_i))} \big|G_\lambda(x, x_i) - \widetilde{K}_\lambda(x, x_i)\big| \le C_{x_i} h^{-d}\big(\tau(n) h^{-d} + h\big).$$

*In particular, the relative error of approximating $G_\lambda(x_i, x_i)$ by the integral (6) satisfies*

$$\frac{\big|G_\lambda(x_i, x_i) - \widetilde{K}_\lambda(x_i, x_i)\big|}{\big|G_\lambda(x_i, x_i)\big|} \le C'_{x_i}\big(\tau(n) h^{-d} + h\big).$$

*Here we may choose $C_{x_i} = C\max\{1, p^{-1/2}(x_i)\}$ and $C'_{x_i} = C\max\{1, p^{1/2-d/(2\alpha)}(x_i)\}\sqrt{p(x_i)}$ for some constant $C$ independent of $(x_i, h, n)$.*

The proof of this theorem relies on a novel Sobolev interpolation inequality that bounds the localized sup-norm via the RKHS norm $\|\cdot\|_{\mathbb{H}}$ plus a localized $L_2$ norm. We remark that the rescaled leverage score $G_\lambda(x_i, x_i)$ and our approximation $\widetilde{K}_\lambda(x_i, x_i)$ would both increase to infinity as $n \to \infty$, and the upper bound for the difference between them would also diverge as shown in the first inequality above; however, the relative error, the quantity of our interest, would shrink. Combining the above with Theorem 2, we can show our approximation leads to an optimal prediction risk in the approximated KRR.

**Theorem 4** (Nyström approximation)**.** *Suppose Assumptions 1, 2 hold for each $x_i, i = 1, \dots, n$. Under the same setting and conditions Theorem 2, if the importance sampling weights are chosen as $q_i = \widetilde{K}_\lambda(x_i, x_i)/\sum_{i=1}^n \widetilde{K}_\lambda(x_i, x_i)$, we have $R_n(\widehat{f}_{L_n}) \le C_3 R_n(\widehat{f})$ with probability at least $1 - 2\rho$.*

## 4 Experiments

In this section, we evaluate our leverage score approximation method on both synthetic and real datasets. The algorithms below are implemented in unoptimized Python code, run with one core of a server CPU (Intel Xeon-Gold 6248 @ 2.50GHZ) on Red Hat 4.8. Specifically, we perform the numerical integration and density estimation as described in Section 9 and 10 in the appendix. Due to the limited space, the complete settings of the experiments below and more supplementary results can be found in Section 7 in the appendix.

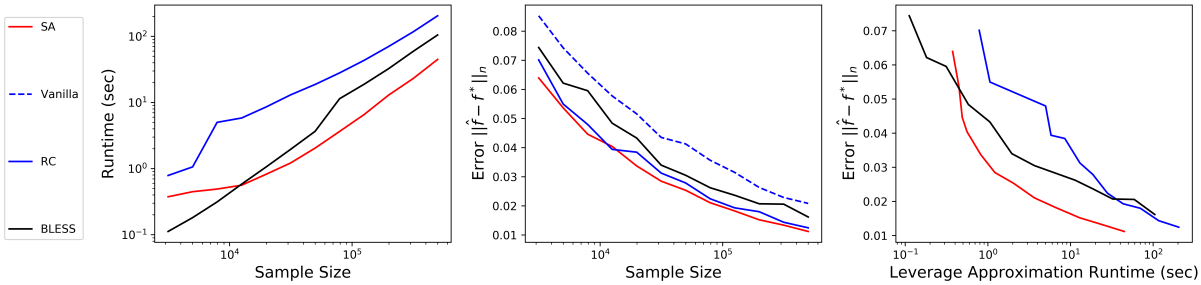


Figure 1: Run time vs. error tradeoff.

### 4.1 Performance in kernel ridge regression

We compare the in-sample prediction error of Nyström methods in KRR, as well as the corresponding leverage approximation time among all the competing algorithms: uniform sampling (hereinafter referred to as "Vanilla"), Recursive-RLS (RC), (Musco and Musco, 2017), Bottom-up Leverage Scores Sampling (BLESS) (Rudi et al., 2018) and our proposed spectral-analysis-based method (SA). (The Monte Carlo approximation for the regularized Christoffel function (Pauwels et al., 2018) in practice reduces to directly computing leverage scores and is thus omitted.) In the experiment, we generate design points $\{x_i\}_{i=1}^n$ with $n \in [2000, 500000]$ from a 3-D bimodal distribution (see Section 7 in the appendix for the definition). We use squared in-sample estimation error $\|\widehat{f} - f^*\|_n^2$ as the evaluation metric. All the results reported in Figure 1 are averaged over 30 replicates. We remark that in the left or the right subplot there is no curve for "Vanilla" method, as this method assumes the leverage scores are uniform and thus takes no time to approximate.

In Figure 1, we can observe "Vanilla" fails to capture the information of the entire design distribution as expected, as with high probability, only few data points from the small mode would be sampled. For RC, BLESS, and our method, although they are all able to capture the non-uniformity, our method has the best runtime versus error trade-off, especially when $n$ is large. Particularly, when $n = 5 \times 10^5$ our method takes 35.8s to approximate the leverage scores, while RC and BLESS respectively take a higher cost—around 94.3s and 167s—due to their higher complexities.

### 4.2 Statistical leverage scores accuracy

We empirically validate that the approximation $\widetilde{K}_\lambda(x_i, x_i)$ approaches the rescaled statistical leverage score $G_\lambda(x_i, x_i)$ as guaranteed by our theory. In particular, we compare the true rescaled leverage and our approximation for samples from one-dimensional (for the ease of visualization) Unif$[0, 1]$, Beta$(15, 2)$, and a bimodal distribution.

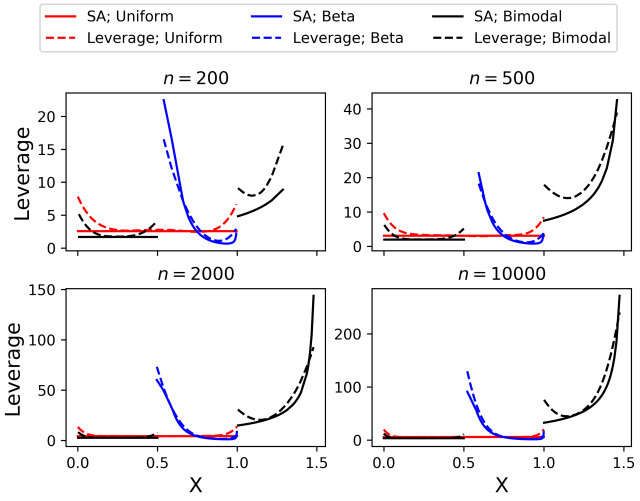


Figure 2: Statistical Leverage Score Approximation

In Figure 2, dotted curves correspond to the rescaled leverage scores, while solid curves correspond to the equivalent kernel approximations. We can observe our method provides good approximations to the rescaled leverage scores across all settings. In particular, Unif$[0, 1]$ is the easiest case (red curves) due to its flat density, which meets Assumption 1 and 2 for almost all design points; while for points with low density, such as those in the smaller cluster of the bimodal distribution and close to the boundary of Beta$(15, 2)$, the absolute

Table 1: Statistical Leverage Score Approximation Accuracy

| | RQC | | | HTRU2 | | | CCPP | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Time | $\bar{r}$ | $5^{th}/95^{th}$ | Time | $\bar{r}$ | $5^{th}/95^{th}$ | Time | $\bar{r}$ | $5^{th}/95^{th}$ |
| SA | 0.40 | 1.01 | 0.87/1.13 | 2.23 | 1.04 | 0.77/1.26 | 0.48 | 1.00 | 0.79/1.21 |
| Vanilla | - | 1.06 | 0.64/1.40 | - | 1.13 | 0.53/1.63 | - | 1.04 | 0.72/1.33 |
| RC | 6.97 | 1.03 | 0.75/1.33 | 2.15 | 1.05 | 0.75/1.27 | 9.21 | 1.02 | 0.82/1.24 |
| Bless | 3.83 | 1.03 | 0.74/1.33 | 1.63 | 1.07 | 0.67/1.32 | 5.25 | 1.02 | 0.81/1.24 |

error tends to be large, due to the leading constant $C_{x_i}$ in the error bound in Theorem 3. Moreover, the relative approximation error has a clear tendency to decrease as the sample size increases, which is consistent with our theory.

We also quantitatively study the accuracy of the leverage scores obtained by different methods in the last section. Each algorithm is tested on `RadiusQueriesCount` (Savva et al., 2018; Anagnostopoulos et al., 2018)(denoted by RQC), `HTRU2` (Lyon et al., 2016), and `CCPP` (Tüfekci, 2014; Kaya and Tüfekci, 2012), the datasets downloaded from the UCI ML Repository (Dua and Graff, 2017). Those datasets contain 10000, 17898, and 9568 data points respectively, which are at the limit of our computational feasibility (it requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space to exactly compute leverage scores). We begin by normalizing the datasets before constructing the kernel matrix using Matérn kernel ($\nu = 0.5$). Each method is then used to approximate the leverage scores $\{\tilde{\ell}_i\}_1^n$. The sampling probability $\tilde{q}_i$ is obtained as $\tilde{\ell}_i/(\sum_1^n \tilde{\ell}_i)$ (also denoting $q_i = \ell_i/(\sum_1^n \ell_i)$). The accuracy of each method is measured by the average of the ratios $\{r_i := \tilde{q}_i/q_i\}_1^n$ (R-ACC). The complete setting for this experiment can be found in Section 7 in the appendix.

In Table 1 we report the runtime, mean R-ACC $\bar{r}$, and the 5th / 95th quantile of R-ACC, averaged over 10 replicates. We notice that, regarding the leverage approximation, our method provides the most accurate leverage approximation (in terms of mean R-ACC), and is more efficient than other methods on the benchmark datasets, which matches the complexity analysis.

### 4.3 Additional empirical results

In Section 10.1 in the appendix, we further provide some empirical results to compare different approximation methods for increasing input dimension $d$. In short, under the certain setting the prediction accuracy of all the methods will greatly deteriorate due to the curse of dimensionality, and the classical Nyström method with uniform sampling will be preferred as leverage-based sampling cannot bring many benefits to the statistical performance.

## 5 Conclusion and future work

We propose a new method to estimate the leverage scores in kernel ridge regression for fast Nyström approximation when a stationary kernel is used. Theoretical results are also provided to guarantee the high accuracy of our estimation. In particular, we show that under the mild conditions the leverage scores induced by a Matérn empirical kernel matrix can be estimated in $\widetilde{\mathcal{O}}(n)$ time, where $n$ is the size of input samples.

A direct further development of our current work is the extension of our theory to other stationary kernels, such as Gaussian kernels and exponential kernels. Other related questions include the performance guarantees when the new leverage estimation method is applied to kernel methods for other machine learning problems, for example, kernel $k$-means and kernel PCA. It will also be interesting to follow the heuristic procedure in our method to analyze other kernel models involving linear smoothers, and seek the possibility to accelerate those models. The results in this work also shed new light on the relevance of the "equivalent kernel" (Yang et al., 2017a) and the regularized Christoffel function (Pauwels et al., 2018) mentioned in Section 1.2.

### Acknowledgements

This work is supported by NSF grant DMS-1810831.

### References

Ahle, T. D., Kapralov, M., Knudsen, J. B., Pagh, R., Velingker, A., Woodruff, D. P., and Zandieh, A. (2020). Oblivious sketching of high-degree polynomial kernels. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 141–160. SIAM.

Alaoui, A. and Mahoney, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783.

Anagnostopoulos, C., Savva, F., and Triantafillou, P. (2018). Scalable aggregation predictive analytics. *Applied Intelligence*, 48(9):2546–2567.

Avron, H., Kapralov, M., Musco, C., Musco, C., Velingker, A., and Zandieh, A. (2017). Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 253–262. JMLR. org.

Backurs, A., Indyk, P., and Wagner, T. (2019). Space and time efficient kernel density estimation in high dimensions. In *Advances in Neural Information Processing Systems*, pages 15773–15782.

Belkin, M. (2018). Approximation beats concentration? an approximation view on inference with smooth radial kernels. In *Conference On Learning Theory*, pages 1348–1361. PMLR.

Berlinet, A. and Thomas-Agnan, C. (2011). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

Calandriello, D., Lazaric, A., and Valko, M. (2017). Distributed adaptive sampling for kernel matrix approximation. In *Artificial Intelligence and Statistics*, pages 1421–1429.

Charikar, M. and Siminelakis, P. (2017). Hashing-based-estimators for kernel density in high dimensions. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1032–1043. IEEE.

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Fukumizu, K. (2008). Elements of positive definite kernel and reproducing kernel hilbert space.

Gittens, A. and Mahoney, M. W. (2016). Revisiting the nyström method for improved large-scale machine learning. *The Journal of Machine Learning Research*, 17(1):3977–4041.

Greengard, L. and Rokhlin, V. (1997). A fast algorithm for particle simulations. *Journal of Computational Physics*, 135(2):280–292.

Greengard, L. and Strain, J. (1991). The fast gauss transform. *SIAM Journal on Scientific and Statistical Computing*, 12(1):79–94.

Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Ivezic, Z., Connolly, A. J., VanderPlas, J. T., and Gray, A. (2014). *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton University Press.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*.

Kaya, H. and Tüfekci, P. (2012). Local and global learning methods for predicting power of a combined gas & steam turbine. In *International Conference on Emerging Trends in Computer and Electronics Engineering*.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

Kumar, S., Mohri, M., and Talwalkar, A. (2009). Sampling techniques for the nystrom method. In *Artificial Intelligence and Statistics*, pages 304–311.

Lyon, R. J., Stappers, B., Cooper, S., Brooke, J., and Knowles, J. (2016). Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1):1104–1123.

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

March, W. B., Xiao, B., and Biros, G. (2015). Askit: Approximate skeletonization kernel-independent treecode in high dimensions. *SIAM Journal on Scientific Computing*, 37(2):A1089–A1110.

Matérn, B. (2013). *Spatial variation*, volume 36. Springer Science & Business Media.

Musco, C. and Musco, C. (2017). Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*, pages 3833–3845.

Mutny, M. and Krause, A. (2018). Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Advances in Neural Information Processing Systems*, pages 9005–9016.

Pauwels, E., Bach, F., and Vert, J.-P. (2018). Relating leverage scores and density using regularized christoffel functions. In *Advances in Neural Information Processing Systems*, pages 1663–1672.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. (2018). On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pages 5672–5682.

Rudi, A., Camoriano, R., and Rosasco, L. (2015). Less is more: Nyström computational regularization. In *Advances in Neural Information Processing Systems*, pages 1657–1665.

Savva, F., Anagnostopoulos, C., and Triantafillou, P. (2018). Explaining aggregates for exploratory analytics. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 478–487. IEEE.

Shawe-Taylor, J., Cristianini, N., et al. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, pages 898–916.

Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140.

Tuo, R., Wang, Y., and Wu, C. (2020). On the improved rates of convergence for mat\'ern-type kernel ridge regression, with application to calibration of computer models. *arXiv preprint arXiv:2001.00152*.

van der Vaart, A. W., van Zanten, J. H., et al. (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, 37(5B):2655–2675.

Van Der Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Walter, G., Blum, J., et al. (1979). Probability density estimation using delta sequences. *the Annals of Statistics*, 7(2):328–340.

Wendland, H. (2004). *Scattered data approximation*, volume 17. Cambridge university press.

Williams, C. K. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688.

Yang, Y., Bhattacharya, A., and Pati, D. (2017a). Frequentist coverage and sup-norm convergence rate in gaussian process regression. *arXiv preprint arXiv:1708.04753*.

Yang, Y., Pilanci, M., Wainwright, M. J., et al. (2017b). Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023.