Accumulations of Projections—A Unified Framework for Random Sketches in Kernel Ridge Regression

Yifan Chen
UIUC
yifanc10@illinois.edu

Yun Yang
UIUC
yy84@illinois.edu

Abstract

Building a sketch of an n-by-n empirical kernel matrix is a common approach to accelerate the computation of many kernel methods. In this paper, we propose a unified framework of constructing sketching methods in kernel ridge regression (KRR), which views the sketching matrix S as an accumulation of m rescaled sub-sampling matrices with independent columns. Our framework incorporates two commonly used sketching methods, sub-sampling sketches (known as the Nyström method) and sub-Gaussian sketches, as special cases with $m = 1$ and $m = \infty$ respectively. Under the new framework, we provide a unified error analysis of sketching approximation and show that our accumulation scheme improves the low accuracy of sub-sampling sketches when certain incoherence characteristic is high, and accelerates the more accurate but computationally heavier sub-Gaussian sketches. By optimally choosing the number m of accumulations, we show that a best trade-off between computational efficiency and statistical accuracy can be achieved. In practice, the sketching method can be as efficiently implemented as the sub-sampling sketches, as only minor extra matrix additions are needed. Our empirical evaluations also demonstrate that the proposed method may attain the accuracy close to sub-Gaussian sketches, while is as efficient as sub-sampling-based sketches.

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

1 Introduction

Kernel methods are widely used in the machine learning community. As a powerful tool, they map data to a higher dimensional feature space and hence can be widely applied to capture nonlinear structures. A well-known representative of kernel methods is the kernel ridge regression (KRR) (Shawe-Taylor et al., 2004; Hastie et al., 2005), which leads to a closed-form estimator and a solid theoretical guarantee on the statistical performance (Stone, 1982; Kanagawa et al., 2018; Tuo et al., 2020). Despite the attractive merits, a practical drawback of KRR is the expensive computational cost. As is common in many other kernel-based machine learning algorithms, solving the KRR with massive data requires manipulating and inverting large matrices usually beyond the capacity of the memory and the time constraint. For example, the evaluation of the empirical kernel matrix K takes $\Theta(n^2)$ time, which already becomes formidable for datasets with hundreds to thousands of samples.

To overcome the limitation, some random projection methods are raised to accelerate the KRR by finding a low-rank approximation to the empirical kernel matrix K. Popular methods include Nyström methods (Williams and Seeger, 2001; Rudi et al., 2015), random features (Rahimi and Recht, 2008), and sketching methods (Woodruff, 2014; Yang et al., 2017; Ahle et al., 2020), the theme of our paper. Actually the sketching method is a general concept that can incorporate many other methods as special cases. One of the most famous examples is the Nyström method, which approximates K by using a subset of "landmark" samples. This procedure to construct an approximation to the kernel matrix is equivalent to the usage of a sub-sampling sketching matrix under our framework. Another example is the sub-Gaussian sketching method.

Different sketching methods have their own advantages and disadvantages. A common observation is that the Nyström method with uniform sampling distribu-

tion needs more "landmark" data points to achieve the same accuracy as a sub-Gaussian sketch (Alaoui and Mahoney, 2015; Yang et al., 2017). To resolve the issue, many works propose data-dependent sampling schemes (Wang and Zhang, 2013; Gittens and Mahoney, 2013; Li et al., 2016) to select the landmarks that improve the approximation of K, Among those weighted sampling schemes statistical leverage-score-based approaches (Alaoui and Mahoney, 2015; Musco and Musco, 2017; Rudi et al., 2018) have attracted most attention due to their superior empirical performances and strong theoretical support. On the other hand, sub-Gaussian sketching methods, although attaining higher statistical accuracy, still suffer from a practically intractable computational cost of at least $\Theta(n^2)$ for evaluating K.

To fully understand the gap in statistical performance between different sketching methods and provide a unified perspective, in this paper we propose a new framework for approximate KRR via randomized sketches. Under this framework, a sketching matrix is taken as an accumulation of m re-scaled sub-sampling matrices with independent columns. The framework incorporates the Nyström method and sub-Gaussian sketches mentioned above as two special cases, by setting $m = 1$ and $m = \infty$ (according to the central limit theorem) respectively. We also analyze the approximation error between the estimator obtained by a sketching method and the original KRR estimator. From the analysis, we find that both the projection dimension d and the number of sub-sampling matrices m contribute to the reduction of the approximation error. Concretely, d addresses the intrinsic difficulty of the KRR problem and is required to be at least $\widetilde{\mathcal{O}}(d_{stat})$ (complexity modulo poly-log term), where d_{stat} is the statistical dimension of the problem; while the number of non-zero elements, or the density of the sketching matrix, indicated by $m \cdot d$, helps mitigate the impact of the high incoherence of the data. More precisely, the incoherence measures the sparsity of the singular vectors of K, and low incoherence means that all coordinates of the singular vectors of K have similar magnitudes instead of few coordinates having dominating values.

Our framework naturally motivates a new method to construct the sketching matrix. Accumulating m sub-sampling matrices by optimally selecting m, the resulting sketching method can achieve the best of two worlds—it achieves high statistical accuracy close to sub-Gaussian sketches, and still retains the high computational efficiency compared to the Nyström methods. The following toy example in Figure 1 illustrates the performance of our method. We run the tests on several synthetic datasets generated from the same distribution while with different sample sizes. (The complete setting of this toy example can be found in the appendix.) We remark here that our method utilizes the sparsity of the sketching matrix to accelerate the training of KRR estimators, and with a medium m it improves the trade-off between accuracy and efficiency.

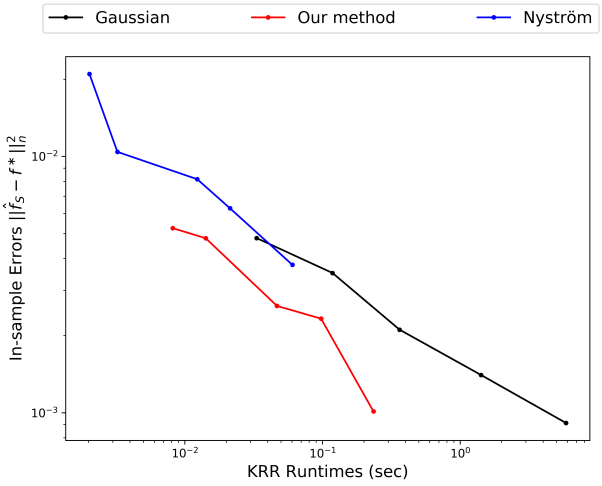

Figure 1: A Toy Example

A similar idea of utilizing sparsity has also appeared in some previous works. Localized sketching (Srinivasa et al., 2020) chooses to use a block-diagonal sketching matrix. Sparse random projections (Achlioptas, 2003; Li et al., 2006) instead randomly set some elements in the projection matrix as zero to increase sparsity. However, there are several important differences between our method and theirs. First, localized sketching assumes the data is partitioned in advance and is mainly designed for distributed or streaming settings. Second, sparse random projections require i.i.d. elements, while our method only requires i.i.d. columns of the sketching matrix—the coordinates in each column are correlated and can follow different distributions. Third, in contrast to sparse random projections, our method exploits some properties of the matrix $K(K + n\lambda I_n)^{-1}$ rather than assume K is a general matrix to be projected. As a result, our method leads to a much sparser sketching matrix than sparse random projections—their matrix density is usually $\sqrt{n}$ times greater than ours.

We also notice that our framework may motivate a broader class of sketching schemes, for example, by using $m \cdot d$ landmarks in the Nyström method, or applies a non-uniform sampling distribution (i.e. leverage-based Nyström method). Our empirical eval-

uations also compare and show superior performance, with another state-of-the-art variant of the Nyström method—Falkon (Rudi et al., 2017).

1.1 Our contributions

In this work, we propose a framework to unify two common and state-of-the-art random sketching methods, and explain the difference in their statistical performances. Specifically, we consider the construction of sketching matrices as the accumulation of m rescaled sub-sampling matrices, which provides a unified perspective to understand how the projection dimension (or size), and the density level of the sketching matrix together influence the approximation error of randomized sketching for KRR, under situations where the incoherence of the data is high and the vanilla Nyström approach fails.

With this framework, we unify existing theoretical results on random sketches and develop a new sketching method to accelerate the training of KRR. This method can be efficiently implemented in practice. Compared to the Nyström method, only some extra parallelizable matrix additions are needed, thereby allowing efficient implementation scalable to massive data. In short, the framework allows the new method to achieve "best of both worlds": it attains the accuracy close to sub-Gaussian sketching while retains the computational efficiency as the sub-sampling sketching.

1.2 Paper outline

The rest of the paper is arranged as follows. In Section 2, we introduce some background on KRR and sketching methods. In Section 3, we illustrate our new framework, provide some theoretical analysis on approximation error, and propose a new sketching method induced by the framework. Finally, in Section 4, we do some numerical experiments to validate our claims in the paper.

2 Preliminaries and notations

Before we present a new framework for random sketches and the related theoretical analysis, we briefly discuss some preliminaries and set up some notations in this section.

2.1 Kernel ridge regression

Consider n pairs of points $\{(x_i, y_i)\}_{i=1}^n$ in $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ is the input (predictor) space and $\mathcal{Y} \subseteq \mathbb{R}$ is the response space. We assume the underlying data generating process is given by the following standard non-parametric regression model,

$$y_i = f^*(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $f^* : \mathcal{X} \to \mathcal{Y}$ is the unknown true regression function, and the noises $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. sub-Gaussian with mean zero and variance σ^2. In this paper, we follow the common regularity assumption that the true regression function $f^* \in \mathbb{H}$, where $\mathbb{H}$ is the reproducing kernel Hilbert space induced by a positive semi-definite kernel function $\mathcal{K}(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ specified beforehand. The true regression function f^*, can be estimated by the *kernel ridge regression* (KRR) (Shawe-Taylor et al., 2004; Hastie et al., 2005), which is cast as the following minimization problem,

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

With the representer theorem, the solution to the optimization problem (1) can be obtained by solving an n-dimensional quadratic program (Kimeldorf and Wahba, 1971). More precisely, after defining K as the n-by-n empirical kernel matrix with $K_{ij} = \mathcal{K}(x_i, x_j)$, $X = (x_1, x_2, \cdots, x_n)^T \in \mathbb{R}^{n\times d_X}$ as the input matrix, $Y \in \mathbb{R}^n$ as the response vector, the solution $\hat{f}_n$ has a closed-form expression as

$$\hat{f}(x) := \mathcal{K}(x, X)\,(K + n\lambda I_n)^{-1}\, Y, \tag{2}$$

where the notation $\mathcal{K}(x, X)$ represents a 1-by-n matrix, whose i-th element is $\mathcal{K}(x, x_i)$.

2.2 Sketching methods

One key issue regarding KRR is that directly solving the problems requires $\mathcal{O}(n^3)$ time complexity and $\mathcal{O}(n^2)$ space complexity, as we need to store and invert an n-by-n matrix $K + n\lambda I_n$. Because of the favorable theoretical properties and the conceptual simplicity of the KRR, researchers have developed a large volume of approximation algorithms to speed up its computation, including sketching methods, the theme of our paper. Sketching methods borrow the idea that random projections can nearly preserve the pairwise distances among column vectors of the empirical kernel matrix K (Arriaga and Vempala, 1999). Concretely, sketching methods utilize the following expression to approximate the empirical kernel matrix K:

$$K_S := KS(S^T K S)^{-1} S^T K,$$

where $S \in \mathbb{R}^{n\times d}$ is the so-called sketching matrix and d is the projection dimension. To obtain the corresponding estimator, we replace K with K_S in the expression (2), and then apply Woodbury matrix identity to calculate an approximate estimator as

$$\begin{aligned}\hat{f}_S(x) :=& K(x, X)S\,(S^T K^2 S \\ &+ n\lambda S^T K S)^{-1} S^T K Y.\end{aligned} \tag{3}$$

As the dimension of the matrix inverted is reduced from n to d, the direct time cost for solving the KRR drops from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 d)$.

A representative of such methods is sub-Gaussian sketching. The name "sub-Gaussian" comes from the property that the columns in a sub-Gaussian sketching matrix are sub-Gaussian random vectors(Pilanci and Wainwright, 2015). In particular, a column s is said to be σ-sub-Gaussian if it's zero-mean, and for any fixed unit vector $u \in \mathbb{S}^{n-1}$, we have

$$\mathbb{P}\{|\langle u, s\rangle| \geq t\} \leq 2\exp(-\frac{nt^2}{2\sigma^2}), \quad \forall t \geq 0.$$

As we implied before, a standard Gaussian sketching matrix with i.i.d. $\mathrm{N}(0,1)$ entries will satisfy the condition and is commonly applied in practice. For sub-Gaussian sketching based on a dense S, such as the standard Gaussian sketching, the $\mathcal{O}(n^2 d)$ computational bottleneck to evaluate the matrix product KS is unavoidable.

To further reduce the $\mathcal{O}(n^2 d)$ time complexity, some researchers turn to sparse sketching matrices. Very sparse random projection (Li et al., 2006) utilizes the random matrices with i.i.d. randomly signed Bernoulli elements (which is a special case of sub-Gaussian sketching); Srinivasa et al. (2020) proposed the localized sketching method that uses a block-diagonal sketching matrix S determined by a partition of the dataset; even more extreme, sub-sampling matrices, in which each column has exactly one non-zero element, are also used in sketching method, known as the Nyström method.

We first give a formal definition of the sub-sampling matrix throughout this paper, which is the basis of many algorithms.

Definition 2.1 (Sub-sampling matrix)**.** *Let P be a discrete distribution which draws $J = j$ with probability $p_j > 0, \forall j \in [n]$. Consider a random matrix $S \in \mathbb{R}^{n\times d}$. If S has i.i.d. columns of the form*

$$\frac{1}{\sqrt{dp_J}} e_J, \quad J \sim P,$$

where e_j is the j-th column of the n-by-n identity matrix I_n, then S is a sub-sampling matrix with sampling distribution P. Specifically, if a random matrix S_R is generated by $S_R = SR_d$, where R_d is a d-by-d diagonal matrix whose entries are i.i.d. Rademacher variables, we call S_R a randomly signed sub-sampling matrix with sampling distribution P.

Remark. *Sub-sampling matrices above are also sub-Gaussian sketching matrices, due to the fact any bounded random variable is sub-Gaussian. However, the sub-Gaussian parameter σ for a random column vector in a sub-sampling matrix grows with n and can be as large as $\mathcal{O}(\sqrt{n})$ for uniform sampling. Therefore we conceptually do not view a sub-sampling matrix as sub-Gaussian (Vershynin, 2010), and prefer to develop distinct theoretical tools for it.*

The classical Nyström method directly uses a sub-sampling matrix with uniform distribution as the sketching matrix S, which reduces the time complexity of computing KS further down to $\mathcal{O}(nd)$ since we only need to store the d chosen columns. Despite the significant reduction of computation complexity, Nyström method is known to suffer from low approximation accuracy (Alaoui and Mahoney, 2015; Yang et al., 2017). We usually need orders of magnitude larger projection dimension d to make the accuracy of sub-sampling sketching comparable to that of sub-Gaussian sketching.

To improve the accuracy of the Nyström method, some papers (Alaoui and Mahoney, 2015; Musco and Musco, 2017; Rudi et al., 2018) suggest to apply a non-uniform sampling distribution P to attain the minimax optimal rate of KRR, via the statistical leverage scores defined below to capture the importance of each sample x_i,

$$\ell_i = \left(K(K + n\lambda I_n)^{-1}\right)_{ii}, \forall i \in [n],$$

where the sampling probability is set as $p_i = \frac{\ell_i}{\sum_{j=1}^n \ell_j}$.

It is also worthy mentioning a related concept, statistical dimension $d_{stat} := \sum_{i=1}^n \ell_i$, motivated by linear regression. Note that d_{stat} can be viewed as an effective rank of the matrix $K(K + n\lambda I_n)^{-1}$, which serves as the theoretical lower bound for any sketching matrix S to maintain the statistical accuracy of the KRR. We will come back shortly in the next section to illustrate this.

3 A unified framework for random sketches

After introducing some notations, in this section, we present our framework, some theoretical results, and our proposed method.

3.1 Introduction to the new framework

We begin by stating an important observation that any n-by-d deterministic dense matrix can be formally taken as the summation of multiple rescaled sub-sampling matrices. Our framework is motivated by the key observation and naturally considers the number of rescaled sub-sampling matrices m as a tunable parameter. This framework induces a new approach to construct the sketching matrix, providing a better trade-off between statistical accuracy and computational ef-

ficiency than existing methods as extreme cases. We use Algorithm 1 below to present the complete setting and procedure to construct a sketching matrix under our framework.

Algorithm 1: Construction of a sketching matrix under the new framework

Input: the size of empirical kernel matrix n, the projection dimension d, the number of sub-sampling matrix m, and the sampling distribution P

Output: An n-by-d sketching matrix S with independent columns

Initialize the matrix $S = 0 \in \mathbb{R}^{n\times d}$;

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

  Sample d numbers $n_1, \cdots, n_d$ from P with replacement;

  Construct an n-by-d rescaled randomly signed sub-sampling matrix $S_{(i)}$, by setting its j-th column as $\frac{r_j}{\sqrt{dmp_{n_j}}} e_{n_j}$ where r_j's are i.i.d. Rademacher variables and e_{n_j} is the n_j-th column of an n-by-n identity matrix;

  Update S as $S + S_{(i)}$;

end

Under this framework, we can recover some important results of prior works introduced in the preliminaries. To show the classical Nyström method with uniform sampling distribution is a special case, we simply set $m = 1$ and P as the uniform distribution over the index set $[n]$. Specifically, the extra random sign due to r_j's will cancel out eventually in forming KS, and the scheme reduces to the Nyström method. As for sub-Gaussian sketching, it can be obtained as the other extreme by letting $m \to \infty$. After element-wisely applying central limit theorem (CLT), S can well approximate a Gaussian sketching matrix, which has sub-Gaussian columns.

3.2 A unified error analysis of randomized KRR

In this subsection, we discuss the influence of m and d on the approximation error for KRR

$$\|\hat{f}_S - \hat{f}_n\|_n^2 := \sum_{i=1}^n |\hat{f}_S(x_i) - \hat{f}_n(x_i)|^2,$$

which is the in-sample squared error between the estimators given by our framework and by the original KRR. Previous works (Yang et al., 2017; Liu et al., 2018) proved that the approximation error is affected by an important property, K-satisfiability, which shows how well the random sketch preserves top eigenspaces of the empirical kernel matrix K. To give the formal definition of K-satisfiability, we need to introduce some heavier notations. Through this paper we denote the n eigenvalues of $\frac{1}{n}K$ as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, and define $d_\delta := \min\{i : \sigma_i \leq \delta\} - 1$, $U_1 \in \mathbb{R}^{n\times d_\delta}$ as the first d_δ columns of U, and $U_2 \in \mathbb{R}^{n\times(n-\delta)}$ as the rest columns. Similarly, Σ_1, and Σ_2 are defined as the diagonal eigenvalue matrices corresponding to U_1 and U_2. We remark here that δ might be unequal to the regularization parameter λ in KRR.

With those notations we define K-satisfiability as follows:

Definition 3.1 (K-satisfiability). *A sketching matrix S is said to be K-satisfiable (for δ) if there exists a constant $c > 0$ such that*

$$\|U_1^T S S^T U_1 - I_{d_\delta}\|_{op} \leq 1/2$$
$$\|S^T U_2 \Sigma_2^{\frac{1}{2}}\|_{op} \leq c\delta^{\frac{1}{2}}$$

The K-satisfiability plays an important role in our main result. Having a sketching matrix with K-satisfiability for $\delta < \lambda$, we can control the squared error $\|\hat{f}_S - \hat{f}_n\|_n^2$ by the following theorem adapted from a previous work (Liu et al., 2018, Theorem 3.8). This theorem assumes some theoretical assumptions of the kernel function $\mathcal{K}$ to control the behavior of the eigenvalues of the kernel function and the empirical kernel matrix (see more details in Section 7 in the appendix). Denoting the spectral expansion of $\mathcal{K}$ as $\mathcal{K}(x, x') = \sum_{i=1}^\infty \mu_i \phi_i(x)\phi_i(x'), \forall x, x' \in \mathcal{X}$, we state the two assumptions made by Liu et al. (2018) as follows.

Assumption 1. *Let $c_K := \sup_{i\geq 1} \|\phi_i\|_{\sup} < \infty$, and $\sup_{k\geq 1} \frac{\sum_{i=k+1}^\infty \mu_i}{k\mu_k} < \infty$.*

Assumption 1 requires the kernel function to have a fast eigenvalue decay rate. Fortunately, many common kernel functions satisfy the assumption. For example, the Matérn kernel with smoothness parameter ν has a decay rate as $\mu_i \asymp i^{-\frac{2\nu+d_{\mathcal{X}}}{d_{\mathcal{X}}}}$ (Bach, 2017), and we can check the rate of this type will satisfy Assumption 1; the Gaussian kernels have an even faster decay rate $\mu_i \asymp \exp(-\gamma \cdot i^c)$, (where γ, c are some constants) and also satisfy the assumption (Liu et al., 2018).

The next assumption has further requirements on the eigenvalue sequence:

Assumption 2. *Let $s_\lambda := \min\{i : \mu_i \leq \lambda\} - 1$. s_λ diverges as $\lambda \to 0$.*

With this assumption, kernels will have a sequence of positive eigenvalues converging to 0. Most infinitely dimensional kernels satisfy this assumption, including the two examples above, Matérn and Gaussian kernels.

Theorem 1 (Approximation error for KRR). *Assume the true regression function f^* lies within the RKHS*

$\mathbb{H}_K$ induced by the kernel function $\mathcal{K}(\cdot,\cdot)$ and the noise vector $\epsilon = Y - f^*(X)$ are i.i.d. sub-Gaussian. $\mathcal{K}$ is further assumed to satisfy Assumption 1 and 2. Let $\delta > 0$ satisfying $\delta < \lambda$. Suppose $d \geq cd_\delta$ for a sufficiently large constant c, and the sketching matrix $S \in \mathbb{R}^{n\times d}$ is K-satisfiable for δ, then with probability approaching 1 we have

$$\|\hat{f}_S - \hat{f}_n\|_n^2 \leq \delta + \frac{d_\delta \delta}{n\lambda} \leq \lambda + \frac{d_\lambda}{n}.$$

When $\delta \ll \lambda$ and $\lambda \to 0$ as $n \to \infty$,

$$\|\hat{f}_S - \hat{f}_n\|_n^2 = o_P(\lambda + \frac{d_\lambda}{n}).$$

Remark. *We briefly discuss how the input dimension d_X affects the error above. Specifically, as the dimension d_X increases, the statistical dimension d_λ will become larger, and therefore the optimal regularization parameter λ, the best possible error ($\mathcal{O}(\lambda)$) (Yang et al., 2017), and the projection dimension d accordingly increase. Taking the Matérn kernel with smoothness parameter ν as an example, its statistical dimension is $\mathcal{O}(n^{\frac{d_X}{2\nu+2d_X}})$ (Yang et al., 2017), and the exponent $\frac{d_X}{2\nu+2d_X}$ increases with d_X. We further remark here that the influence of an increasing dimension p is universal, not specific to our proposed method, and all prospective methods have to use a larger projection dimension d. This is because even the optimal approximation, the truncated SVD, needs to have rank at least d_λ.*

The result guarantees a small estimation error $\|\hat{f}_S - f^*\|_n^2$. The error upper bound $\lambda + \frac{d_\lambda}{n}$ is comparable to the estimation error $\|\hat{f}_n - f^*\|_n^2$ (Yang et al., 2017). Hence the theorem suggests that the key is to verify the K-satisfiability of S. The next theorem provides conditions to guarantee the K-satisfiability to hold with high probability. The complete statement and the proof are deferred to the appendix.

Theorem 2 (Conditions on d and m)**.** *Let δ be a regularization parameter which might be different than λ, and $\rho \in (0, 1/2)$. Define the incoherence*

$$M := \max\left\{\max_i \frac{\|\tilde{\psi}_i\|^2}{p_i}, \max_i \frac{\|\psi_i\|^2 - \|\tilde{\psi}_i\|^2}{p_i}\right\},$$

where $\{p_i\}_{i=1}^n$ is the sub-sampling probability and $\tilde{\psi}_i$ is the sub-vector of the first d_δ elements in the i-th column ψ_i of $\Psi_\delta = [\Sigma(\Sigma + n\delta I_n)]^{-1/2} U^T$. Given the assumptions in Theorem 1, there exist two constants $c_1, c_2 > 0$ such that if a sketching matrix $S \in \mathbb{R}^{n\times d}$ constructed by Algorithm 1 with m accumulations satisfies

$$\begin{aligned} d &\geq c_1 d_\delta \log^2(\frac{n}{\rho}) \\ md &\geq c_2 M \log^3(\frac{n}{\rho}), \end{aligned} \tag{4}$$

then with probability at least $1-\rho$, the sketching matrix S satisfies K-satisfiability with δ as the regularization parameter.

Theorem 2 reveals that d needs at least to be d_δ to address the intrinsic difficulty of KRR, while it is unnecessary to have a larger d as m can make up for the gap between d and M. It also gives the reason why sub-Gaussian sketches ($m = \infty$) outperform uniform sub-sampling sketches with the same size d, in terms of the statistical performance.

We further point out when the classical Nyström method with uniform sampling distribution is applied, the incoherence M can be of the same magnitude as n. For example, suppose the kernel function $\mathcal{K}$ used is compactly supported, and we intentionally manipulate the input X to have two clusters far away from each other, one dense cluster of size d_δ and the other sparse cluster of size $n - d_\delta$. As the kernel function is compactly supported, $\mathcal{K}(x_i, x_j) = 0$ for x_i and x_j from two clusters respectively. The empirical kernel matrix K thus becomes block-diagonal, and the eigendecomposition of $K = U\Sigma U^T$ shares the same property that after rearranging the order of columns, U becomes block-diagonal as well. Since the small cluster is denser than the other, assuming the large cluster is extremely sparse, it is possible to assign the greatest d_δ eigenvalues all to the dense cluster, and we can verify (in this certain case)

$$M \geq \frac{\|\tilde{\psi}_1\|^2}{1/n} \geq \frac{n}{2}\|\tilde{u}_1\|^2 = \frac{n}{2},$$

where $\tilde{u}_1$ is the truncated vector of the first d_δ elements in U's first column u_1. The second inequality holds as $\frac{\sigma_i}{\sigma_i+\delta} \geq \frac{1}{2}\, \forall i \leq d_\delta$; the final equation holds as U is block diagonal and $\|\tilde{u}_1\|^2 = \|u_1\|^2 = 1$.

This example implies that unbalanced multimodal data brings high incoherence to the KRR problem, which explains why the classical Nyström method fails in certain cases (Yang et al., 2017).

Remark. Theorem 2 recovers the statistical leverage scores based Nyström methods. When the subsampling probability p_i, as suggested by previous literature (Alaoui and Mahoney, 2015; Musco and Musco, 2017; Rudi et al., 2018), is proportional to the statistical leverage score ℓ_i, we have

$$\begin{aligned} M &\leq \max_i \frac{\|\psi_i\|^2}{p_i} = \|\Psi_\delta\|_F^2 \\ &= d_{stat} := \sum_{i=1}^n \frac{\sigma_i}{\sigma_i + \delta} \sim d_\delta. \end{aligned}$$

The last equivalence relation means that d_δ and d_{stat} have the same magnitude with high probability (Liu

et al., 2018, Lemma 3.1). Fixing $m = 1$, the special sampling distribution substantially reduces the incoherence M, and thus mitigates the lower bound requirement on d. Theorem 2 also explains why sub-Gaussian sketches can attain the same statistical accuracy as the original KRR without explicit correction.

3.3 A new sketching method

Our unified framework naturally introduces a tunable parameter m for Algorithm 1 to construct a best random sketch in practice. Specifically, after picking $d = \widetilde{\mathcal{O}}(d_\delta), m = \widetilde{\mathcal{O}}(\frac{M}{d_\lambda})$, we first generate md samples from the given sampling distribution P, then construct m rescaled sub-sampling matrix $S_{(i)}$'s, and finally sum them up to obtain the sketching matrix S.

At first sight, we may lose the computational advantage of sketching via a sub-sampling matrix, and may suffer from the high computational cost of using a sub-Gaussian sketch. Fortunately, similar to the very sparse random projection (Li et al., 2006), the sketch of the empirical kernel matrix can be efficiently constructed by few extra matrix addition steps. For the approximate KRR estimator (3), we observe that the main bottleneck is the computation of KS and S^TKS. To address the issue, we can utilize the equations

$$KS = \sum_{i=1}^{m} KS_{(i)}, \quad S^TKS = \sum_{i=1}^{m} S_{(i)}^T(KS),$$

to reduce the time cost of computing KS and S^TKS from $\mathcal{O}(n^2d)$ and $\mathcal{O}(nd^2)$, to $\mathcal{O}(nmd)$ and $\mathcal{O}(md^2)$ respectively. The overall time cost to solve KRR problems then becomes $\mathcal{O}(nd^2)$, assuming no other approximation methods are used. Compared to the Nyström method, we improve the statistical performance at the cost of some extra matrix additions. The extra cost can be further reduced under a parallel computation architecture, since matrix addition is highly parallelizable on a GPU.

According to Theorem 2 our proposed method above has some advantages over commonly used schemes implied by Theorem 2, namely using Nyström method while increasing the size of the sketch to $\widetilde{\mathcal{O}}(M)$, or using the leverage-based Nyström method. To make the comparison fair, throughout the subsection we consider a vanilla scheme that uses a sub-sampling sketch with dimension $n \times md$, where m, d are the parameters used in our method. The first benefit of our method is, by picking the smallest possible d, the sketching matrix S, as well as the matrix S^TK^2S, maintains a small size. As the computation of the matrix S^TK^2S is the bottleneck, the total runtime of the vanilla scheme is roughly m^2 times more than ours. We remark here that the factor m^2 is not always negligible, as our example in the last section shows that M can be much larger than d_δ. Besides the higher time cost, $m > 1$ also implies the vanilla method needs to invert a bigger matrix while solving the KRR problem, which may deteriorate the numerical stability. As for the leverage-based Nyström method, we point out it will cost $\mathcal{O}(n^3)$ time to exactly compute the statistical leverage scores. Recently some approximation methods have been proposed to estimate statistical leverage scores. However, the multiplicative error between the leverage scores and their approximation demands a larger projection dimension d.

Now we discuss and compare some other methods for accelerating the training of KRR. Since the development of approximation methods to train KRR is irrelevant to the theme of our work, we only introduce one of the most representative models, Falkon (Rudi et al., 2017). This method combines several different algorithmic principles. Besides sketching methods, it also incorporates early stopping and preconditioning to reduce the KRR training time to $\mathcal{O}(nd_{stat}\log n)$. To attain this complexity, we need to additionally assume an access to the q-approximate leverage scores $\hat{\ell}_i(\lambda)$'s, satisfying $q^{-1}\ell_i(\lambda) \le \hat{\ell}_i(\lambda) \le q\ell_i(\lambda), \forall i \in [n]$, as the runtime of Falkon also depends on the quality of the sketching matrix.

In this case, the total runtime for the vanilla scheme with Falkon is $\mathcal{O}(nmd\log n)$, while the time cost for us is $\mathcal{O}(nmd + nd\log n)$. Note that there is no m in the last term as the runtime of Falkon is related to the dimension of S. In addition to the time complexity comparison, we add a closing remark that considering Falkon is an iterative method with multiple matrix inversions in each step, our method may benefit Falkon by reducing the matrix size from md to d. Some numerical experiments in the appendix further demonstrate our method will have the best comprehensive performance among all the candidate approaches.

4 Experiments

In this section, we test our results based on a range of datasets. All the experiments below are implemented in unoptimized Python code, run with four cores of a server CPU (Intel Xeon-Gold 6248 @ 2.50GHZ) on Red Hat 4.8. Due to the limited space of the paper, the complete experiment settings and more details are deferred to the appendix.

In Section 4.1, we corroborate our theories by comparing the approximation error for different choices of m on synthetic datasets. The result in Figure 2 demonstrates that for the classical Nyström method with uniform sampling, a medium m is enough to attain the high statistical accuracy close to sub-Gaussian

sketches. This is remarkable as the sketching matrix constructed by our method is much sparser than sub-Gaussian sketches, and only requires a fraction of the computational resources. In Section 4.2, we compare the prediction accuracy of the estimators obtained by Gaussian sketching (Yang et al., 2017), very sparse random projections (Li et al., 2006), the Nyström method with leverage scores approximated by BLESS (Rudi et al., 2018), and our accumulation method using uniform sampling distribution on a certain real dataset. The results are displayed in Figure 3. We also consider the usage of Falkon (Rudi et al., 2017), and provide some more experiment results in the appendix to show our method still attain the optimal trade-off between statistical accuracy and computational efficiency.

4.1 Approximation error

In this experiment, we compare the approximation error $\|\hat{f}_S - \hat{f}_n\|_n^2$ among the KRR estimators obtained by the sketching matrices with different m, including the Gaussian sketching as an instance of $m = \infty$. In particular, we ran the experiment on a bimodal distribution over $\mathbb{R}^3$. The bimodal distribution has two components: with probability $\frac{n}{n+n^\gamma}$ generating a Unif$[0,1]^3$; and with probability $\frac{n^\gamma}{n+n^\gamma}$ generating a random variable with pdf $\prod_{j=1}^3(5-2x_j)$ for $x_j \in [2, 2.5]^3$, where n is the sample size varying from $n = 1,000$ to $8,000$ and $\gamma = 0.6$. In addition, by cross validation the Gaussian kernel with bandwidth $\sigma = 1.5n^{-\frac{1}{7}}$ is chosen, and the regularization parameter of the KRR λ is set as $0.5n^{-\frac{4}{7}}$. The true regression function f^* we use is

$$f^*(x) = 1.6|(x-0.4)(x-0.6)| - x(x-1)(x-2) - 0.5,$$

and i.i.d. noises follow $\mathcal{N}(0, 0.25)$. We use a uniform sub-sampling distribution for all the applicable methods. The results reported in Figure 2 are averaged over 30 replicates, and the error bars are computed as the standard error of the corresponding average values. In Figure 2, we actually take the Nyström method ($m = 1$, the blue solid line) and the Gaussian sketching method ($m = \infty$, the black dashed line) as two benchmarks. The estimation error $\|\hat{f}_n - f^*\|_n^2$ of the original KRR (the black solid line) is also provided in the figure as a reference. As expected, with the same projection dimension d, Gaussian sketches attain the approximation error orders of magnitude lower than the Nyström method. However, as implied by Theorem 2, with a medium $m \leq 32$ the approximation error can be reduced to a similar scale as the Gaussian sketch. The phenomenon validates the effectiveness of Theorem 2, and suggests that in practice a medium m is enough to make our sketching method as accurate as Gaussian sketching.

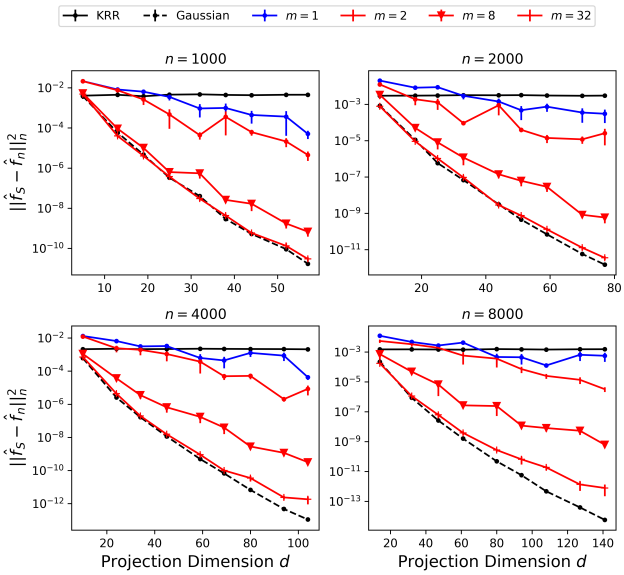

Figure 2: Approximation Errors of Different Methods

4.2 Accuracy versus efficiency trade-off

To show the improvement on the comprehensive performance, we evaluate our method on `RadiusQueriesAggregation` (Savva et al., 2018; Anagnostopoulos et al., 2018)(denoted by RQA), a dataset downloaded from the UCI ML Repository (Dua and Graff, 2017). This dataset contains 200000 data points and 4 features. Due to the space limit, here we only present a partial result, and the rest parts will be provided in the appendix.

To show the evolving trend, we set a sequence of sample sizes beforehand and in each round run the experiment on a subset of the whole data points with the given sample size n. The testing errors are estimated on a random subset (20% of the original dataset) which is not used in the training. We begin by normalizing the features to have variance 1 in the randomly drawn dataset, before obtaining the empirical kernel matrix using Matérn kernel (the smoothness parameter $\nu = 1.5$). The regularization parameter λ of KRR is $0.9 \cdot n^{-7/11}$. We set the projection dimension as $\lfloor 1.5 \cdot n^{4/11} \rfloor$ for all sketching methods. The candidate methods include the Gaussian sketching method, very sparse random projection (Li et al., 2006), the Nyström method with BLESS (Rudi et al., 2018), and

our accumulation method with $m = 4$. All the results reported below in Figure 3 are averaged over 30 replicates, and the error bars are computed as before.

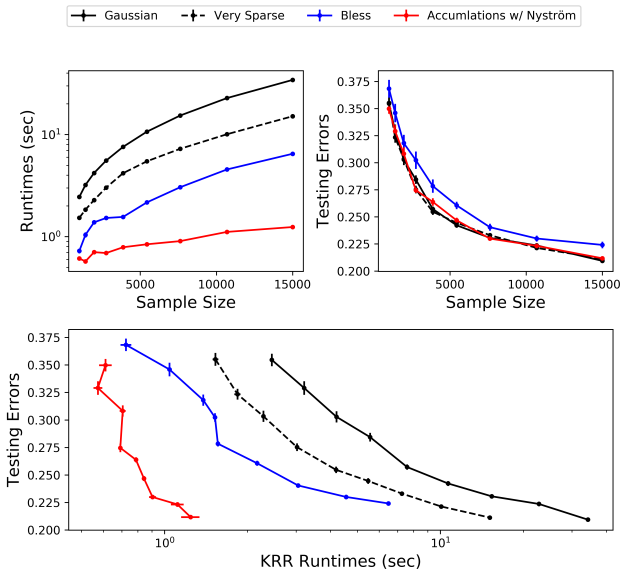

Figure 3: Trade-off between Accuracy and Efficiency

In Figure 3 we compare the comprehensive performance of those methods. The comprehensive performance is composed of two factors, statistical accuracy, and computational efficiency. As we can see in the last subplot of Figure 3, our method attains the best trade-off between in-sample error and runtime among all the candidate methods. Taking a closer look, we can even observe that our method achieves the advantages of the two extremes in our framework: it can attain an accuracy similar to Gaussian sketching, and enjoy the low runtime of the same order as the Nyström method. As for very sparse random projection, since it is designed for a general matrix and does not utilize the leverage information, its performance falls somewhere in between the above-described methods. We ascribe the remarkable performance of our method to the proper sparsity of the sketching matrix and the high efficiency of the matrix addition.

4.3 Additional empirical results

In the appendix, we verify the effect of our method on the other two datasets `PPGasEmission` (KAYA et al., 2019) and `CASP` (Dua and Graff, 2017), and also confirm the usage of Falkon (Rudi et al., 2017) does not affect the main claim above. Basically, those experiments demonstrate that in practice, a medium m substantially improves the accuracy of the classical Nyström method with uniform sub-sampling distribution, and the extra cost is much lower compared to other advanced methods.

5 Conclusion and future work

We have introduced a unified framework for randomized sketches in kernel ridge regression. This framework unifies two common methods, sub-Gaussian and sub-sampling sketching, by introducing an accumulation parameter m. Our theoretical results state that a medium m improves the prediction performance of the approximate KRR, while still maintaining a low projection dimension even when the sub-sampling probability p_i's are not optimal. The empirical experiments complement our theory by showing that our proposed method with medium m and simple uniform sampling scheme can achieve high accuracy close to sub-Gaussian sketching, and be as efficient as sub-sampling-based sketching.

One possible extension of our work is applying the proposed sketching method to approximate matrix multiplication and how the approximation error translates when the new sketching method is utilized to approximate some classical machine learning models, such as k-means and PCA. Another direction of future research is the extension of our method to broader classes of positive semidefinite kernels, for example graph kernels (Vishwanathan et al., 2010) and string kernels (Vishwanathan and Smola, 2002).

Acknowledgements

This work is supported by NSF grant DMS-1810831.

References

Achlioptas, D. (2003). Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687.

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.

Arriaga, R. I. and Vempala, S. (1999). An algorithmic theory of learning: Robust concepts and random projection. In *40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039)*, pages 616–623. IEEE.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751.

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

Gittens, A. and Mahoney, M. (2013). Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pages 567–575. PMLR.

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.

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., TÜFEKCİ, P., and UZUN, E. (2019). Predicting CO and NOxemissions from gas turbines: novel data and abenchmark PEMS. *TURKISH JOURNAL OF ELECTRICAL ENGINEERING & COMPUTER SCIENCES*, 27(6):4783–4796.

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

Li, C., Jegelka, S., and Sra, S. (2016). Fast dpp sampling for nystr\" om with application to kernel methods. *arXiv preprint arXiv:1603.06052*.

Li, P., Hastie, T. J., and Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296.

Liu, M., Shang, Z., and Cheng, G. (2018). Nonparametric testing under random projection. *arXiv preprint arXiv:1802.06308*.

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

Pilanci, M. and Wainwright, M. J. (2015). Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115.

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.

Rudi, A., Carratino, L., and Rosasco, L. (2017). Falkon: An optimal large scale kernel method. In *Advances in Neural Information Processing Systems*, pages 3888–3898.

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.

Srinivasa, R. S., Davenport, M. A., and Romberg, J. (2020). Localized sketching for matrix multiplication and ridge regression. *arXiv preprint arXiv:2003.09097*.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053.

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*.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vishwanathan, S. and Smola, A. J. (2002). Fast kernels for string and tree matching. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, pages 585–592.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242.

Wang, S. and Zhang, Z. (2013). Improving cur matrix decomposition and the nyström approximation via adaptive sampling. *The Journal of Machine Learning Research*, 14(1):2729–2769.

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.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*.

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