Fast Randomized Kernel Methods With Statistical Guarantees

Ahmed El Alaoui * Michael W. Mahoney †

Abstract

One approach to improving the running time of kernel-based machine learning methods is to build a small sketch of the input and use it in lieu of the full kernel matrix in the machine learning task of interest. Here, we describe a version of this approach that comes with running time guarantees as well as improved guarantees on its statistical performance. By extending the notion of *statistical leverage scores* to the setting of kernel ridge regression, our main statistical result is to identify an importance sampling distribution that reduces the size of the sketch (i.e., the required number of columns to be sampled) to the *effective dimensionality* of the problem. This quantity is often much smaller than previous bounds that depend on the *maximal degrees of freedom*. Our main algorithmic result is to present a fast algorithm to compute approximations to these scores. This algorithm runs in time that is linear in the number of samples—more precisely, the running time is $O(np^2)$, where the parameter p depends only on the trace of the kernel matrix and the regularization parameter—and it can be applied to the matrix of feature vectors, without having to form the full kernel matrix. This is obtained via a variant of length-squared sampling that we adapt to the kernel setting in a way that is of independent interest. Lastly, we provide empirical results illustrating our theory, and we discuss how this new notion of the statistical leverage of a data point captures in a fine way the difficulty of the original statistical learning problem.

1 Introduction

We consider the low-rank approximation of positive semidefinite matrices that arise in machine learning and data analysis, with an emphasis on obtaining good statistical guarantees. This is of interest primarily in connection with kernel-based methods. Recent work in this area has focused on one or the other of two very different perspectives: an *algorithmic perspective*, where the focus is on running time issues and worst-case quality-of-approximation guarantees, given a fixed input matrix; and a *statistical perspective*, where the goal is to obtain good inferential properties, under some hypothesized model, by using the low-rank approximation in place of the full kernel matrix. The recent results of Gittens and Mahoney [1] provide the strongest example of the former, and the recent results of Bach [2] are an excellent example of the latter. In this paper, we combine ideas from these two lines of work in order to obtain a fast randomized kernel method with statistical guarantees that are improved relative to the state-of-the-art.

To understand our approach, recall that several papers have established the crucial importance—from the algorithmic perspective—of the *statistical leverage scores*, as they capture important structural non-uniformities of the input matrix and they can be used to obtain very sharp worst-case approximation guarantees. See, e.g., work on CUR matrix decompositions [3, 4], work on the fast approximation of the statistical leverage scores [5], and the recent review of Randomized

*Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720. Email: elalaoui@eecs.berkeley.edu

†International Computer Science Institute and Department of Statistics, University of California at Berkeley, Berkeley, CA 94720. Email: mmahoney@stat.berkeley.edu

Linear Algebra [6] for more details. Here, we simply note that, when restricted to an $n \times n$ positive semidefinite matrix K and a rank parameter k, the *statistical leverage scores relative to the best rank-k approximation to K*, call them $\ell_i (= \ell_i(k))$, for $i \in \{1, \ldots, n\}$, are the diagonal elements of the projection matrix onto the best rank-k approximation to K. That is, $\ell_i = \mathrm{diag}(K_k K_k^\dagger)_i$, where K_k is the best rank k approximation to K and where $K_k^\dagger$ is the Moore-Penrose inverse of K_k. The recent work by Gittens and Mahoney [1] showed that qualitatively-improved worst-case bounds for the low-rank approximation of positive semidefinite matrices could be obtained in one of two related ways: either compute (with the fast algorithm of [5]) approximations to the leverage scores, and use those approximations as an importance sampling distribution in a random sampling algorithm; or rotate (with a Gaussian-based or Hadamard-based random projection) to a random basis where those scores are approximately uniformized, and sample randomly in that rotated basis.

In this paper, we extend these ideas, and we show that—from a statistical perspective—we are able to obtain a low-rank approximation that comes with improved statistical guarantees by using a variant of this notion of statistical leverage. In particular, we improve the recent bounds of Bach [2], which provide the first known statistical convergence result when substituting the kernel matrix by its low-rank approximation. To understand the connection, recall that a key component of Bach's approach is the quantity $d_{\mathrm{mof}} = n \| \mathrm{diag}(\, K(K + n\lambda I)^{-1}) \|_\infty$, which he calls the *maximal degrees of freedom*. Bach's main result is that by constructing a low-rank approximation of the original kernel matrix by sampling uniformly at random $p = \Theta(d_{\mathrm{mof}}/\epsilon)$ columns, i.e., performing the vanilla Nyström method, and then by using this low-rank approximation in a prediction task, the statistical performance is within a factor of $1 + \epsilon$ of the performance when the entire kernel matrix is used.

Here, we show that this uniform sampling is suboptimal. We do so by sampling with respect to a variant of the statistical leverage scores which we call the *λ-ridge leverage scores*. These are defined in Definition 1 below, and we show that we can obtain similar $1+\epsilon$ statistical performance guarantees by sampling only $\Theta(d_{\mathrm{eff}}/\epsilon)$ columns, where $d_{\mathrm{eff}} = \mathsf{Tr}(K(K + n\lambda I)^{-1}) < d_{\mathrm{mof}}$. The quantity d_{eff} is called the *effective dimensionality* of the learning problem, and it can be interpreted as the implicit number of parameters in this nonparametric setting [7, 8]. In addition, we show that these λ-ridge leverage scores can be approximated quickly and without forming the full kernel matrix. In particular, this can be accomplished in $O(np^2)$ time, where p is a low-rank parameter related to the λ-ridge leverage scores. We accomplish this by performing a variant of length-squared sampling on the original feature vectors, and then using the matrix inversion lemma to measure how "outlying" each data point is, which is a technique of independent interest.

Given the interplay between the algorithmic and statistical approaches that we have adopted and exploited, we expect that our results and insights will be useful much more generally. As an example of this, we can directly compare the Nyström sampling method to a related divide-and-conquer approach, thereby answering an open problem of Zhang et al. [7]. Recall that the Zhang et al. divide-and-conquer method consists of dividing the dataset $\{(x_i, y_i)\}_{i=1}^n$ into m random partitions of equal size, computing estimators on each partition in parallel, and then averaging the estimators. They prove the minimax optimality of their estimator, although their multiplicative constants are suboptimal; and, in terms of the number of kernel evaluations, their method requires $m \times (n/m)^2$, with m in the order of n/d_{eff}^2, which gives a total number of $O(nd_{\mathrm{eff}}^2)$ evaluations. They noticed that the scaling of their estimator was *not* directly comparable to that of the Nyström sampling method (which was proven by Bach to require $O(nd_{\mathrm{mof}})$ evaluations, if the sampling is uniform [2]), and they left it as an open problem to determine which if either method is fundamentally better than the other. Using our Theorem 3, we are able to put both results on a common ground for comparison. Indeed, the estimator obtained by our *non-uniform* Nyström

sampling requires only $O(nd_{\text{eff}})$ kernel evaluations (improving both $O(nd_{\text{eff}}^2)$ and $O(nd_{\text{mof}})$), and it obtains the same bound on the statistical predictive performance as in [2]. In this sense, our result combines "the best of both worlds," by having the reduced sample complexity of [7] and the sharp approximation bound of [2].

2 Preliminaries and notation

Let $\{(x_i, y_i)\}_{i=1}^n$ be n pairs of points in $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the response space. The kernel-based learning problem can be cast as the following minimization problem:

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

where $\mathcal{F}$ is a reproducing kernel Hilbert space and $\ell : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$ is a loss function. We denote by $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the positive definite kernel corresponding to $\mathcal{F}$ and by $\phi : \mathcal{X} \to \mathcal{F}$ a corresponding feature map. That is, $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$ for every $x, x' \in \mathcal{X}$. The representer theorem [9, 10] allows us to reduce Problem (1) to a finite-dimensional optimization problem, in which case Problem (1) boils down to finding the vector $\alpha \in \mathbb{R}^n$ that solves

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

where $K_{ij} = k(x_i, x_j)$. We let $U \Sigma U^\top$ be the eigenvalue decomposition of K, with U an orthogonal matrix and $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. The underlying data model is

$$y_i = f^*(x_i) + \sigma^2 \xi_i \quad i = 1, \cdots, n,$$

with $f^* \in \mathcal{F}$, $(x_i)_{i=1}^n$ a deterministic sequence and ξ_i i.i.d. standard normal random variables. Here, we consider ℓ to be the squared loss, in which case we will be interested in the mean squared error as a measure of statistical risk: for any estimator $\hat{f}$, let

$$\mathcal{R}(\hat{f}) := \frac{1}{n} \mathsf{E}_\xi \|\hat{f} - f^*\|_2^2 \tag{3}$$

be the risk function of $\hat{f}$, where E_ξ denotes the expectation under the randomness induced by ξ. In this setting, the problem is called *Kernel Ridge Regression* (KRR). The solution to Problem (2) is $\alpha = (K + n\lambda I)^{-1} y$, and the estimate of f^* at any training point x_i is given by $\hat{f}(x_i) = (K\alpha)_i$. We will use $\hat{f}_K$ as a shorthand for the vector $(\hat{f}(x_i))_{i=1}^n \in \mathbb{R}^n$ when the matrix K is used as a kernel matrix. This notation will be used accordingly for other kernel matrices (e.g., $\hat{f}_L$ for a matrix L). Recall that the risk of the estimator $\hat{f}_K$ can then be decomposed into a bias and variance term:

$$\begin{aligned}
\mathcal{R}(\hat{f}_K) &= \frac{1}{n} \mathsf{E}_\xi \|K(K + n\lambda I)^{-1}(f^* + \xi) - f^*\|_2^2 \\
&= \frac{1}{n} \|(K(K + n\lambda I)^{-1} - I) f^*\|_2^2 \\
&\qquad + \frac{1}{n} \mathsf{E}_\xi \|K(K + n\lambda I)^{-1} \xi\|_2^2 \\
&= n\lambda^2 \|(K + n\lambda I)^{-1} f^*\|_2^2 + \frac{\sigma^2}{n} \mathsf{Tr}(K^2 (K + n\lambda I)^{-2}) \\
&:= \qquad \text{bias}(K)^2 \qquad + \qquad \text{variance}(K).
\end{aligned} \tag{4}$$

From these expressions, we see that the bias is matrix-decreasing and the variance is matrix-increasing in K. This will be a useful fact in later proofs and discussions.

Solving Problem (2), either by a direct method or by an iterative optimization algorithm needs at least quadratic and often cubic running time in n, which is prohibitive in the large-scale setting. The so-called Nyström method approximates the solution to Problem (2) by substituting K with a low-rank approximation to K. In practice, this approximation is often not only fast to construct, but the resulting learning problem is also often easier to solve [11, 12, 2, 1]. The method operates as follows. A small number of columns $K_1, \cdots, K_p$ are randomly sampled from K. If we let $C = [K_1, \cdots, K_p] \in \mathbb{R}^{n \times p}$ denote the matrix containing the sampled columns and $W \in \mathbb{R}^{p \times p}$ the overlap between C and $C^\top$ in K, then the Nyström approximation of K is the matrix

$$L = CW^\dagger C^\top.$$

More generally, if we let $S \in R^{n \times p}$ be an arbitrary *sketching matrix*, i.e., a tall and skinny matrix that, when left-multiplied by K, produces a "sketch" of K that preserves some desirable properties of K, then the Nyström approximation associated with S is

$$L = KS(S^\top KS)^\dagger S^\top K. \tag{5}$$

For instance, for random sampling algorithms, S would contain a non-zero entry at position (i, j) if the i-th column of K is chosen at the j-th trial of the sampling process. Alternatively, S could also be a random projection matrix; or S could be constructed with a some other (perhaps deterministic) method, as long as it satisfies some structural properties, e.g., those given below or those given in [1], depending on the application [6, 1, 4, 3].

3 Our main results

In this section, we present our main algorithmic and statistical results in the statistical prediction context of estimating f^* by approximating the solution to Problem (2). To accomplish this, we proceed by revisiting and improving upon three prior results. These improvements are of independent interest, and we will use them in the analysis of our main results. The first (in Section 3.1) is on the behavior of the bias of $\hat{f}_L$, when L is constructed using a general sketching matrix S; the second (in Section 3.2) is a concentration bound for approximating matrix multiplication when the rank-one components of the product are sampled nonuniformly; and the third (in Section 3.3) is an extension of the definition of the statistical leverage scores to the context of kernel ridge regression. By combining and extending these three improvements, we are able to obtain our main results: a strong statistical statement (in Section 3.4) and algorithmic statement (in Section 3.5) on the behavior of the Nyström method if one is allowed to sample nonuniformly.

3.1 A basic structural result

We begin by establishing a basic "structural" result that upper-bounds the bias of the estimator constructed using the approximation L of Eqn. (5) in place of the full kernel K. The proof of this theorem may be found in Appendix A.

Theorem 1 *Let $S \in \mathbb{R}^{n \times p}$ be a sketching matrix and L the corresponding Nyström approximation. For $\gamma > 0$, let $\Phi = \Sigma(\Sigma + n\gamma I)^{-1}$. If the sketching matrix S satisfies*

$$\lambda_{\max}\Big(\Phi - \Phi^{1/2} U^\top S S^\top U \Phi^{1/2}\Big) \leq t,$$

for $t \in (0,1)$ and $\gamma \leq (1-t)\lambda$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue, then

$$bias(L) \leq \left(1 - \frac{\gamma/\lambda}{1-t}\right)^{-1} bias(K). \tag{6}$$

In the special case where S is a matrix representing the process of sampling uniformly-at-random, i.e., where S contains one non-zero entry equal to $1/\sqrt{pn}$ in every column, where p is the number of sampled columns, this result and its proof can be found in Appendix B.2 of [2]. Our more general result follows easily with a slight change of notation by considering any arbitrary sketching matrix S.

We emphasize that this structural result is a deterministic statement: it only depends on the properties of the input data, and holds for *any* sketching matrix S that satisfies certain conditions (stated in the theorem). Thus, any randomness enters only via S, and the randomness in the construction of S is "decoupled" from the rest of the analysis. We highlight this fact since this two-step approach was adopted previously [3, 1] and since it offers a clear way to improve our current results: a better construction of S—whether deterministic or randomized, whether sampling-based or projection-based, etc.—satisfying the data-dependent conditions stated in the theorem would immediately lead to downstream algorithmic and statistical improvements for kernel ridge regression in this setting.

Our subsequent statistical analysis (for Theorem 3 below) will unfold in two steps. First, assuming that the sketching matrix S satisfies the conditions stated in Theorem 1, we will have $\mathcal{R}(\hat{f}_L) \lesssim \mathcal{R}(\hat{f}_K)$. (This will follow since, from the proof of Theorem 1, $L \preceq K$, and thus by monotonicity the variance in Eqn. (4) will decrease; but the bias in Eqn. (4) will increase, when replacing K by L; and thus to obtain a bound on $\mathcal{R}(\hat{f}_L)$, it suffices to control the bias term.) Second, a matrix concentration bound will be used to show that an appropriate random construction of S corresponding to a random sampling process satisfies these conditions. To that end, we next state the concentration result that is the source of our improvement (in Section 3.2), and we then define a notion of statistical leverage scores (in Section 3.3) that will be used as an importance sampling distribution by the sampling process.

3.2 A concentration bound on approximate matrix multiplication

Next, we establish our result for approximating matrix products of the form $\Psi\Psi^\top$, when a small number of columns from the $n \times m$ matrix Ψ (and thus rows of $\Psi^\top$) are sampled to form the approximate product $\Psi_I \Psi_I^\top$, where Ψ_I is a matrix that contains the chosen columns. The proof of this theorem, which uses a matrix Bernstein inequality,[1] may be found in Appendix B.

Theorem 2 *Let n, m be positive integers. Consider a matrix $\Psi \in \mathbb{R}^{n\times m}$ and denote by ψ_i the i^{th} column of Ψ. Let $p \leq m$ and $I = \{i_1, \cdots, i_p\}$ be a subset of $\{1, \cdots, m\}$ formed by p elements chosen randomly with replacement, according to the distribution*

$$\forall i \in \{1, \cdots, m\} \quad \mathsf{Pr}(\textit{choosing } i) = p_i \geq \beta \frac{\|\psi_i\|_2^2}{\|\Psi\|_F^2}, \tag{7}$$

[1] The proof of Theorem 2 uses one of the state-of-the-art bounds on matrix concentration, but it is one among many other related bounds in the literature. While it constitutes a base for our improvement, it is possible that a concentration bound more tailored to the problem might yield still sharper results.

for some $\beta \in (0,1]$. Let $S \in \mathbb{R}^{n\times p}$ be a sketching matrix such that $S_{ij} = 1/\sqrt{p \cdot p_{i_j}}$ only if $i = i_j$ and 0 elsewhere. Then

$$\begin{aligned}&\mathsf{Pr}\Big(\lambda_{\max}\big(\Psi\Psi^\top - \Psi SS^\top\Psi^\top\big) \geq t\Big)\\ &\quad\leq n\exp\left(\frac{-pt^2/2}{\lambda_{\max}(\Psi\Psi^\top)(\|\Psi\|_F^2/\beta + t/3)}\right).\end{aligned} \tag{8}$$

Remark. This result holds for arbitrary Ψ, but we will apply it to $\Psi = \Phi^{1/2}U^\top$, where $\Phi = \Sigma(\Sigma + n\gamma I)^{-1}$, and in conjunction with Theorem 1 it can be used to prove our main statistical result in Theorem 3 below. Notice that $\Psi^\top$ is a scaled version of the eigenvectors, with a scaling given by the square root of the diagonal matrix Φ. This can be considered as a "soft projection" or "soft truncation" matrix that smoothly selects the top part of the spectrum of K. The setting of Gittens and Mahoney [1], in which Φ is a 0-1 diagonal matrix (and thus which is more like a Tikhonov smoothing than a ridge smoothing) is the closest analog to the use of Φ in our setting.

Remark. It is known that $p_i = \frac{\|\psi_i\|_2^2}{\|\Psi\|_F^2}$ is the optimal sampling distribution in terms of minimizing the expected error $\mathsf{E}\|\Psi\Psi^\top - \Psi SS^\top\Psi^\top\|_F^2$; see [13]. Importantly, Theorem 2 exhibits a robustness property by allowing the chosen sampling distribution to be different than the optimal one by a factor β.[2] The effect of the sub-optimality of such a distribution is reflected in the upper bound in Eqn. (8) by multiplying the squared Frobenius norm of Ψ by a factor $1/\beta$. In particular, if the sampling distribution is chosen to be uniform, i.e., $p_i = 1/m$, $\forall i$, then the value of β for which Eqn. (7) is tight is

$$\frac{\|\Psi\|_F^2}{m\max_i \|\psi_i\|_2^2},$$

in which case we recover the previous concentration result proven by Bach [2].

3.3 An extended notion of statistical leverage

Here, we introduce a notion of statistical leverage that is tailored to the ridge regression problem and that will be used in our analysis; we call these scores the *λ-ridge leverage scores*.

Definition 1 *For $\lambda > 0$, the λ-ridge leverage scores associated with the kernel matrix K and the parameter λ are*

$$\forall i \in \{1,\cdots,n\}, \quad l_i(\lambda) = diag(K(K+n\lambda I)^{-1})_i. \tag{9}$$

In words, $l_i = l_i(\lambda)$ is the i^{th} diagonal entry of the matrix $K(K+n\lambda I)^{-1}$. Observe that $l_i(\lambda) = \sum_{j=1}^n \frac{\sigma_j}{\sigma_j+n\lambda}U_{ij}^2$; and, moreover, that for $\Psi = \Sigma^{1/2}(\Sigma+n\lambda I)^{-1/2}U^\top$, we have that $\|\psi_i\|_2^2 = l_i(\lambda)$. These formulae provide naïve ways to compute the λ-ridge leverage scores exactly; but below we will present a way to compute approximations to them much more quickly.

The quantities $(l_i(\lambda))_{i=1,\cdots,n}$ are analogs of the so-called *leverage scores* in the statistical literature [14]. These quantities are classically defined as the diagonal elements of the projection matrix onto the input matrix (or equivalently as the Euclidean norms of the rows of the left singular vector matrix U of the input matrix), and they have been used in regression diagnostics

[2] In their work [13], Drineas et al. have a comparable robustness statement for controlling the expected error. Our result is a robust quantification of the tail probability of the error, which is a much stronger statement.

for outlier detection [14]. More recently, they have been used as an algorithmic tool in Randomized Linear Algebra, as they provide an importance sampling distribution for constructing random sketches for low-rank approximation [3, 4, 1] as well as least-squares regression [15] when the input matrix is tall and skinny. (In the case where the input matrix is square and full-rank, this definition is uninteresting, as the row norms of U are all equal to 1.) Recently, Gittens and Mahoney [1] used a truncated version of these scores (the so-called *leverage scores relative to the best rank-k space*, which equal $\ell_i = \mathrm{diag}(K_k K_k^\dagger)_i$, where K_k is the best rank-k approximation to K) to obtain the best algorithmic results known to date on low-rank approximation of positive semidefinite matrices. Definition 1 is a weighted version of these leverage scores, where the weights depend on the spectrum of K and a regularization parameter λ. In this sense, they provide an interpolation between the leverage scores relative to the best rank-k space and the classical (tall-and-skinny) leverage scores, where the parameter λ plays the role of a rank or interpolation parameter.

We should point out that Bach's maximal degrees of freedom d_{mof} is to the λ-ridge leverage scores what the *coherence* μ is to leverage scores relative to the best rank-k space $\ell_i = \ell_i(k)$ used by Gittens and Mahoney: $d_{\mathrm{mof}}/n = \max_i l_i(\lambda)$, and $\mu/n = \max_i \ell_i(k)$, where μ is the matrix coherence. We should also point out that both sum (or average) to a natural capacity parameter: $\sum_{i=1}^n \ell_i = k$, and $\sum_{i=1}^n l_i = d_{\mathrm{eff}}$. Definition 1 provides a relevant notion of leverage in the statistical setting of kernel ridge regression, and it is a natural counterpart in the statistical prediction context of the algorithmic notion of leverage relative to the best rank-k space.

3.4 Main statistical result: an error bound on approximate kernel ridge regression

We are now ready to present our main statistical result, which is a theorem that establishes an improved performance guarantee on the use of the Nyström method in the context of kernel ridge regression. It is improved in the sense that the sufficient number of columns that need be sampled in order to incur very little loss in the prediction performance is lower than the previous result of Bach [2]. Our improvement comes from using a more data-sensitive importance sampling distribution (depending on the λ-ridge leverage scores) when choosing the columns of K during the construction of L. The proof of this theorem may be found in Appendix C.

Theorem 3 *Let $\lambda, \rho > 0$, $\epsilon \in (0, 1/2)$, and L be a Nyström approximation of K by choosing p columns randomly with replacement according to a probability distribution $(p_i)_{1 \le i \le n}$ such that*

$$\forall i \in \{1, \cdots, n\}, \quad p_i \geq \beta \cdot l_i(\lambda\epsilon) / \sum_{i=1}^n l_i(\lambda\epsilon),$$

for some $\beta \in (0, 1]$. If

$$p \geq 8\Big(\frac{d_{\mathit{eff}}}{\beta} + \frac{1}{6}\Big) \log(\frac{n}{\rho}),$$

with $d_{\mathit{eff}} = \sum_{i=1}^n l_i(\lambda\epsilon) = \mathsf{Tr}(K(K + n\lambda\epsilon I)^{-1})$, then

$$\mathcal{R}(\hat{f}_L) \leq (1 - 2\epsilon)^{-2} \mathcal{R}(\hat{f}_K)$$

with probability at least $1 - \rho$, where $(l_i)_i$ are introduced in Definition 1 and $\mathcal{R}$ is defined in Eqn. (3).

In words, Theorem 3 asserts that substituting the kernel matrix K by a Nyström approximation of rank p in the kernel ridge regression problem induces an arbitrarily small prediction

loss, provided that p scales linearly with the effective dimensionality d_{eff}.[3] The leverage-based sampling appears to be crucial for obtaining this improved dependency, as the λ-ridge leverage scores provide information on which columns—and hence which data points—capture most of the difficulty of the estimation problem. If the columns are sampled uniformly, then a much worse lower bound on p that depends on d_{mof} is obtained [2].

Note that, the smaller the target accuracy ϵ, the higher d_{eff}, and in general the more uniform the sampling distribution $(l_i(\lambda\epsilon))_i$ becomes; in the limit $\epsilon \to 0$, p is in the order of n and the leverage scores become uniform, in which case the method is essentially equivalent to using the entire matrix K. Note also that, if the sampling distribution $(p_i)_i$ is a factor β away from optimal, then a slight oversampling (i.e., increase p by $1/\beta$) achieves the same performance. In this sense, our result shows robustness to the exact sampling distribution. This property is very important from an implementation point of view, as the error bounds still hold when only an approximation of the leverage scores is available.

3.5 Main algorithmic result: a fast approximation to the λ-ridge leverage scores

We are now ready to present our main algorithmic result. Although the λ-ridge leverage scores can be naïvely computed exactly by performing a full SVD computation, this exact computation is as costly as solving the original Problem (2). Here we present an algorithm that computes approximations to these scores much more quickly.[4] By the robustness properties described above, these approximations can be used in place of the exact λ-leverage scores. We start with a description of our algorithm.

Algorithm:

- **Inputs**: data points $(x_i)_{i=1}^n$, probability vector $(p_i)_{i=1}^n$, sampling parameter $p \in \{1, 2, \cdots\}$, $\lambda > 0$, and $\epsilon \in (0, 1/2)$.
- **Output**: $(\tilde{l}_i)_{1 \le i \le n}$, ϵ-approximations to $(l_i(\lambda))_{1 \le i \le n}$.

1. Sample p data points from $(x_i)_{i=1}^n$ with replacement with probabilities $(p_i)_{i=1}^n$.
2. Compute the corresponding columns $K_1, \cdots, K_p$ of the kernel matrix.
3. Construct $C = [K_1, \cdots, K_p] \in \mathbb{R}^{n \times p}$ and $W \in \mathbb{R}^{p \times p}$ as presented in Section 2.
4. Construct $B \in \mathbb{R}^{n \times p}$ such that $BB^\top = CW^\dagger C^\top$.
5. For every $i \in \{1, \cdots, n\}$, set

$$\tilde{l}_i = B_i^\top (B^\top B + n\lambda I)^{-1} B_i \tag{10}$$

where B_i is the i-th row of B, and return it.

[3] Note that d_{eff} depends on the precision parameter ϵ, which is absent in the classical definition of the effective dimensionality [8, 7, 2]. However, the following bound holds: $d_{\text{eff}} \le \frac{1}{\epsilon}\mathsf{Tr}(K(K + n\lambda I)^{-1})$.

[4] This is similar to the spectral and Frobenius approximation algorithms of Drineas et al. [5] that provide very fine approximations to the leverage scores (for tall and skinny matrices) and the leverage scores relative to the best rank-k space (for general matrices) in roughly the time it takes to implement and apply a random projection to the matrix.

Running time. The running time of this algorithm is dominated by steps 4 and 5. Indeed, constructing B can be done using a Cholesky factorization on W and then a multiplication of C by the inverse of the obtained Cholesky factor, which yields a running time of $O(p^3 + np^2)$. Computing the approximate leverage scores $(\tilde{l}_i)_{1\leq i\leq n}$ in step 5 also runs in $O(p^3 + np^2)$. Thus, for $p \ll n$, the overall algorithm runs in $O(np^2)$.

Remark. Note that the formula given in Eqn. (10) only involves matrices and vectors of size p, i.e., everything is computed in the smaller dimension p, and the fact that this yields a correct approximation relies on the matrix inversion lemma. Note also that only the relevant columns of K are computed, and the algorithm never has to form the entire kernel matrix. This is a substantial improvement over earlier models, e.g., that of [1], that are formulated in such a way that the entire matrix K must be written down in memory. Note finally that the improved running time is obtained by considering the construction given in Eqn. (10); this is quite different than typical methods that have been used to approximate leverage scores that instead involve approximating the subspace [5].

Quality of approximation. We now give both additive and multiplicative error bounds on the approximation quality of this algorithm. The proof of this theorem may be found in Appendix D.

Theorem 4 *Let $\epsilon \in (0, 1/2)$, $\rho \in (0, 1)$ and $\lambda > 0$. Let L be a Nyström approximation of K by choosing p columns at random with probabilities $p_i = K_{ii}/\mathsf{Tr}(K)$, $i = 1, \cdots, n$. If*

$$p \geq 8\Big(\frac{\mathsf{Tr}(K)}{n\lambda\epsilon} + \frac{1}{6}\Big)\log(\frac{n}{\rho})$$

then we have $\forall i \in \{1, \cdots, n\}$

$$\textit{(additive error)} \quad l_i(\lambda) - 2\epsilon \leq \tilde{l}_i \leq l_i(\lambda)$$

and

$$\textit{(multiplicative error)} \ \Big(\frac{\sigma_n - n\lambda\epsilon}{\sigma_n + n\lambda\epsilon}\Big)l_i(\lambda) \leq \tilde{l}_i \leq l_i(\lambda)$$

with probability at least $1 - \rho$.

We conclude our main results with several remarks.

Remark. In words, Theorem 4 states that if the columns of K are sampled proportionally to K_{ii}, then $\Theta(\frac{\mathsf{Tr}(K)}{n\lambda})$ is a sufficient number of samples to obtain bounds of the form stated. Recall that $K_{ii} = \|\phi(x_i)\|^2_{\mathcal{F}}$, and so our procedure is akin to sampling according to the length-squared of the original data vectors. This has been extensively used in different contexts in Randomized Linear Algebra in general [16, 13, 17, 6] and in the low-rank approximation of Gram matrices in particular [18]. Conversely, if $p_i = \frac{1}{n}$, for $i = 1, \cdots, n$, then we obtain the same additive and multiplicative error bounds above, but the number of sampled columns must scale like $\Theta(d_{\mathrm{mof}}) = \Theta(n \max_i l_i(\lambda))$ instead.

Remark. In virtue of Theorem 3, the above multiplicative error bound—although weaker than the additive error bound since it depends on the minimum eigenvalue of K—directly provides a bound on the prediction error when $(\tilde{l}_i)_i$ are used as a sampling distribution.

Remark. Due to how λ is defined in Eqn. (1), the n in the denominator should *not* be of concern: $n\lambda$ should be thought of as a "rescaled" regularization parameter λ'. In some settings, the λ that yields the best generalization error scales like $1/\sqrt{n}$, and hence $p = \Theta(\mathsf{Tr}(K)/\sqrt{n})$ would be sufficient. On the other hand, if the columns are sampled uniformly, one would need $p = \Theta(d_{\mathrm{mof}})$.

Remark. From the proof of Theorem 3, it is easy to see that when using the algorithm above to approximate $(l_i(\lambda))$ within an error of ϵ, the optimal sampling distribution to be used is $(l_i(\lambda\epsilon))$. Given the robustness property with respect to the sampling distribution certified by Theorem 2, using other upper bound on the $\lambda\epsilon$-ridge scores will suffice, albeit requiring a larger number of sampled columns. A particularly simple such upper bound is given by the diagonal entries of the kernel matrix:

$$l_i(\lambda\epsilon) = \sum_{j=1}^{n} \frac{\sigma_j}{\sigma_j + n\lambda\epsilon} U_{ij}^2 \leq \sum_{j=1}^{n} \frac{\sigma_j}{n\lambda\epsilon} U_{ij}^2 = \frac{1}{n\lambda\epsilon} K_{ii}.$$

Hence, when $p_i = K_{ii}/\mathsf{Tr}(K)$ is used as the importance sampling distribution in this algorithm, we have $p_i \geq \beta\ l_i(\lambda\epsilon)/\sum_{i=1}^{n} l_i(\lambda\epsilon)$, with $\beta = n\lambda\epsilon d_{\text{eff}}/\mathsf{Tr}(K)$.

Remark. Subsequent to the completion of the preliminary version of this paper, we learned of a paper that also used a structure of the general form of Eqn. (10) to compute approximations to the leverage scores of a tall and skinny matrix [19], thus providing a very different method than that used in [5] to approximate the leverage scores of a tall and skinny matrix. It is an interesting open problem to determine whether the iterative technique used in [19] can lead to improvements in our results and/or in the approximation of the leverage scores relative to the best rank-k space of a matrix.

4 Empirical observations

We illustrate our results with several datasets, one synthetic regression problem from [2] and several real datasets, including the *Pumadyn* family, consisting of three datasets *pumadyn-32fm*, *pumadyn-32fh* and *pumadyn-32nh*[5], and the *Gas Sensor Array Drift Dataset* from the UCI database[6]. Rather than providing a complete evaluation of our method, our goal here is two-fold: first, to illustrate the difference between d_{eff} and d_{mof} in simplified settings; and second, to provide an example illustrating the importance of the λ-ridge leverage scores.

kernel	dataset	n	nb. feat	σ	λ	d_{eff}	d_{mof}
Bern.	Synthetic	500	-	-	$1e{-}6$	24	500
Linear	Gas2	1244	128	-	1	126	1244
	Gas3	1586	128	-	$1e{-}3$	125	1586
	Pum-32fm	2000	32	-	1	30	2000
	Pum-32fh	2000	32	-	1	30	2000
	Pum-32nh	2000	32	-	1	32	2000
RBF	Gas2	1244	-	1	$4.5e{-}4$	1135	1244
	Gas3	1586	-	1	$5e{-}4$	1450	1586
	Pum-32fm	2000	-	5	0.5	142	1897
	Pum-32fh	2000	-	5	$5e{-}2$	747	1989
	Pum-32nh	2000	-	5	$1.3e{-}2$	1337	1997

Table 1: Parameters and quantities of interest for the different datasets and using different kernels: the synthetic dataset using the Bernoulli kernel (denoted by Synthetic); the Gas Sensor Array Drift Dataset (batches 2 and 3, denoted by Gas2 and Gas3); and the Pumadyn datasets (Pum-32fm, Pum-32fh, Pum-32nh) using both linear and RBF kernels.

[5] http://www.cs.toronto.edu/~delve/data/pumadyn/desc.html

[6] https://archive.ics.uci.edu/ml/datasets/Gas+Sensor+Array+Drift+Dataset

For the former goal, we did the following. For all the datasets, we determined λ by cross validation, and we computed the effective dimensionality d_{eff} and the maximal degrees of freedom d_{mof}. Table 1 summarizes these results. Observe how the effective dimensionality d_{eff} is always smaller and typically much smaller than the maximum degrees of freedom d_{mof}.

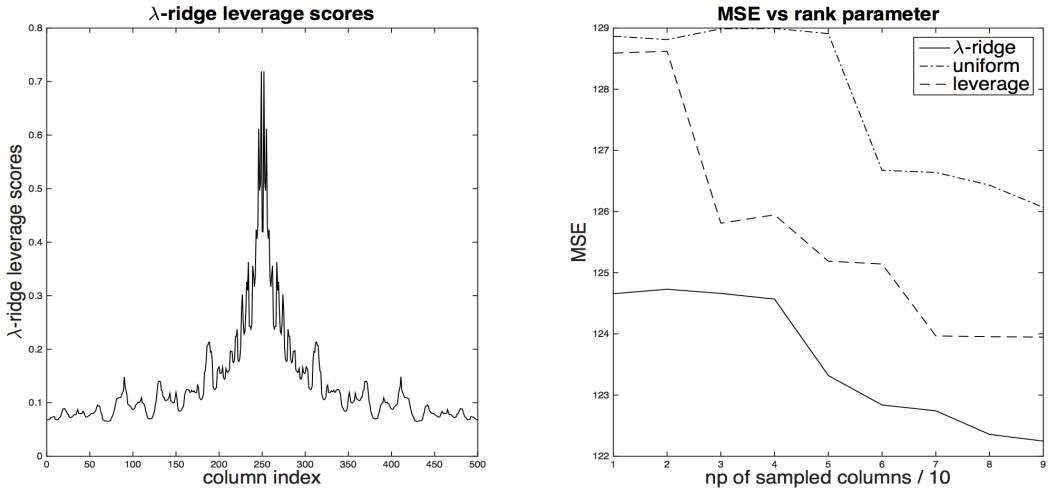

Figure 1: The λ-ridge leverage scores for the synthetic Bernoulli polynomial kernel dataset described in the text (left); and the MSE risk vs. the number of sampled columns used to construct the Nyström approximation for different sampling methods (right).

For the latter goal, we worked with the synthetic dataset. The synthetic dataset consists of a regression problem on the unit interval $\mathcal{X} = [0, 1]$ where, given a signal sequence $(x_i)_{1 \le i \le n}$ and a noise sequence $(\epsilon_i)_{1 \le i \le n}$, we observe the sequence

$$y_i = f(x_i) + \sigma^2 \epsilon_i, \quad i \in \{1, \cdots, n\}.$$

The function f belongs to the RKHS $\mathcal{F}$ generated by the kernel

$$k(x, y) = \frac{1}{(2\beta)!} B_{2\beta}(x - y - \lfloor x - y \rfloor),$$

where $B_{2\beta}$ is the 2β-th Bernoulli polynomial [2]. One important feature of this regression problem is the distribution of the points (x_i) on the unit interval $\mathcal{X}$. If they are spread uniformly (or nearly uniformly) over the unit interval, then the λ-ridge leverage scores $(l_i(\lambda))$ are uniform (or nearly uniform), for every $\lambda > 0$. In this case, uniform column sampling is optimal (or nearly optimal). In fact, if $x_i = \frac{i-1}{n}$ for $i = 1, \cdots, n$, then the kernel matrix K is a circulant matrix [2], in which case one can prove that the λ-ridge leverage scores are constant. On the other hand, if the data points are distributed nonuniformly in the unit interval, then the λ-ridge leverage scores are nonuniform, and using a nonuniform importance sampling distribution is beneficial. See, e.g., Figure 1. For this figure, the data points $x_i \in (0, 1)$ have been generated with a distribution symmetric about $\frac{1}{2}$, having a high density on the borders of the unit interval $(0, 1)$ and a low density at the center of the unit interval. The number of observations is $n = 500$. On Figure 1, we can see that there are extremely few data points with very high leverage, and those that do have high leverage correspond to points in the the unit interval that are underrepresented in the data sample (i.e., the region close to the center of the interval, since that is where one has the lowest density of observations). The λ-ridge leverage scores—and our fast algorithm for computing approximations to them—are able to capture the importance (i.e., influence or leverage, as in the analysis of outliers [20, 14]) of these data points, thus providing a way to detect them.

5 Conclusion

We have shown that the sampling complexity of CUR-based or Nyström-based methods for solving kernel ridge regression can be reduced to the effective dimensionality of the problem, hence bridging and improving upon different previous attempts that established weaker forms of this result. This was achieved by combining and improving upon results that have emerged in recent years from two different perspectives on low-rank matrix approximation; and it involved defining a natural analog to the notion of statistical leverage scores and using them as an importance sampling distribution. Importantly, we also present a computationally tractable way to approximate these scores, i.e., it runs in $O(np^2)$ time, with p depending only on the trace of the kernel matrix and the regularization parameter. Importantly, this algorithm does not require the formation of the full kernel matrix. One natural unanswered question is whether it is possible to reduce further the sampling complexity, i.e., is the effective dimensionality a lower bound? In addition, as pointed out by previous work [2], it is unclear whether the same or similar results can be obtained beyond the squared loss, e.g., for logistic regression or support vector regression. Finally, it is of interest to understand better the connection between the recent algorithm of [19] and our use of Eqn. (10) in our main algorithm.

Acknowledgements

We thank Yuchen Zhang for pointing out the connection to his work. We thank Francis Bach for stimulating discussions.

References

[1] A. Gittens and M. W. Mahoney. Revisiting the Nyström method for improved large-scale machine learning. Technical report, 2013. Preprint: arXiv:1303.1849 (2013).

[2] F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 185–209, 2013.

[3] P. Drineas, M.W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

[4] M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.

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

[6] M. W. Mahoney. *Randomized algorithms for matrices and data*. Foundations and Trends in Machine Learning. NOW Publishers, Boston, 2011. Also available at: arXiv:1104.5557.

[7] Y. Zhang, J. Duchi, and M. Wainwright. Divide and conquer kernel ridge regression. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 592–617, 2013.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2003.

[9] G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2):495–502, 1970.

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

[11] S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

[12] C.K.I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Annual Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference*, pages 682–688, 2001.

[13] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices I: Approximating matrix multiplication. *SIAM Journal on Computing*, 36:132–157, 2006.

[14] S. Chatterjee and A. S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

[15] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2010.

[16] A. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, 51(6):1025–1041, 2004.

[17] P. Drineas, R. Kannan, and M. W. Mahoney. Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36:158–183, 2006.

[18] P. Drineas and M. W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

[19] M. B. Cohen, Y. T. Lee, C. Musco, C. Musco, R. Peng, and A. Sidford. Uniform sampling for matrix approximation. Technical report, 2014. Preprint: arXiv:1408.5099 (2014).

[20] P.F. Velleman and R.E. Welsch. Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242, 1981.

[21] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A Proof of Theorem 1

The structure of this proof is due to Bach [2]; we simply re-derive the same analysis for the case of a more general sketching matrix S.

For kernel ridge regression, the bias of the estimator $\hat{f}_K$ can be expressed as

$$\begin{aligned}\mathrm{bias}(K)^2 &= n\lambda^2 \|(K+n\lambda I)^{-1} f^*\|^2 \\ &= n\lambda^2 {f^*}^\top (K+n\lambda I)^{-2} f^*.\end{aligned}$$

For $\gamma > 0$, we consider the regularized approximation

$$L_\gamma = KS(S^\top KS + n\gamma I)^{-1} S^\top K,$$

with $S \in \mathbb{R}^{n\times p}$ the sketching matrix. With $K = U\Sigma U^\top$, $R = \Sigma^{1/2} U^\top S$ and $\bar{L}_\gamma = R(R^\top R + n\gamma I)^{-1} R^\top$, we have

$$L_\gamma = U\Sigma^{1/2} \bar{L}_\gamma \Sigma^{1/2} U^\top.$$

Due to the matrix inversion lemma, we have

$$\begin{aligned}\bar{L}_\gamma &= RR^\top (RR^\top + n\gamma I)^{-1} \\ &= I - n\gamma (RR^\top + n\gamma I)^{-1} \\ &= I - n\gamma (\Sigma + n\gamma I + RR^\top - \Sigma)^{-1} \\ &= I - n\gamma (\Sigma + n\gamma I)^{-1/2} (I - D)^{-1} (\Sigma + n\gamma I)^{-1/2},\end{aligned}$$

with

$$\begin{aligned}D &= (\Sigma + n\gamma I)^{-1/2} (\Sigma - RR^\top)(\Sigma + n\gamma I)^{-1/2} \\ &= \Phi - \Phi^{1/2} U^\top S S^\top U \Phi^{1/2},\end{aligned}$$

and $\Phi = \Sigma(\Sigma + n\gamma I)^{-1}$. This shows that for any $\gamma \geq 0$

$$L_\gamma \preceq L \preceq K. \tag{11}$$

Now if $\lambda_{\max}(D) \leq t$ for $t \in (0,1)$, then

$$I - \bar{L}_\gamma \preceq \frac{n\gamma}{1-t} (\Sigma + n\gamma I)^{-1},$$

which implies

$$K - L_\gamma \preceq \frac{n\gamma}{1-t} K(K + n\gamma)^{-1} \preceq \frac{n\gamma}{1-t} I.$$

Then, assuming $\frac{\gamma/\lambda}{1-t} \leq 1$, the previous inequality implies

$$\begin{aligned}\big(L_\gamma + n\lambda I\big)^{-1} &\preceq \big(K - \frac{n\gamma}{1-t} I + n\lambda I\big)^{-1} \\ &= \big(K + n\lambda(1 - \frac{\gamma/\lambda}{1-t}) I\big)^{-1} \\ &\preceq \left(1 - \frac{\gamma/\lambda}{1-t}\right)^{-1} \big(K + n\lambda I\big)^{-1}.\end{aligned}$$

Hence

$$\mathrm{bias}(L_\gamma)^2 \leq \left(1 - \frac{\gamma/\lambda}{1-t}\right)^{-2} \mathrm{bias}(K)^2.$$

From $L_\gamma \preceq L$, and the fact that $\mathrm{bias}(K)$ is matrix-decreasing in K, we obtain the desired result.

B Proof of Theorem 2

This proof uses the matrix Bernstein inequality; we will use the variant provided by Theorem 6.1.1 of [21]).

Theorem 5 *Consider a sequence (X_k) of independent random symmetric matrices with dimension d. Assume that $\mathsf{E}(X_k) = 0$, $\lambda_{\max}(X_k) \leq R$, and let $Y = \sum_k X_k$. Furthermore, assume that there exists $\sigma > 0$ such that $\|\mathsf{E}(Y^2)\|_2 \leq \sigma^2$. Then*

$$\mathsf{Pr}\big(\lambda_{\max}(Y) \geq t\big) \leq d \exp\Big(\frac{-t^2/2}{\sigma^2 + Rt/3}\Big).$$

To use this result, we will exhibit the sequence (X_k) and Y in our case. We have

$$\Psi\Psi^\top = \sum_{i=1}^m \psi_i \psi_i^\top$$

and

$$\Psi S S^\top \Psi^\top = \frac{1}{p} \sum_{i \in I} \frac{1}{p_i} \psi_i \psi_i^\top = \frac{1}{p} \sum_{i=1}^m \sum_{k=1}^p \frac{1}{p_i} z_{ik} \psi_i \psi_i^\top,$$

where z_{ik} are iid $0-1$ random variables for $i \in \{1, \cdots, m\}$ and $k \in \{1, \cdots, p\}$ with $\mathsf{Pr}(z_{ik} = 1) = p_i$. Let $Y = \Psi\Psi^\top - \Psi S S^\top \Psi^\top$, then

$$Y = \frac{1}{p} \sum_{k=1}^p \sum_{i=1}^m (1 - \frac{z_{ik}}{p_i}) \psi_i \psi_i^\top.$$

We choose X_k to be $\frac{1}{p} \sum_{i=1}^m (1 - \frac{z_{ik}}{p_i}) \psi_i \psi_i^\top$ for every $k \in \{1, \cdots, p\}$. Now we verify the assumptions of Theorem 5. The matrices X_k inherit independence from the random variables z_{ik}, and we have $\mathsf{E}(X_k) = 0$, and $\lambda_{\max}(X_k) \leq \frac{1}{p} \lambda_{\max}(\sum_{i=1}^m \psi_i \psi_i^\top) = \frac{1}{p} \lambda_{\max}(\Psi\Psi^\top)$. Now we control the spectral norm of the second moment of Y. Again with $\mathsf{E}(X_k) = 0$ we have $\mathsf{E}(Y^2) = \sum_{k,k'=1}^p \mathsf{E}(X_k X_{k'}) = \sum_{k=1}^p \mathsf{E}(X_k^2)$. And for $k \in \{1, \cdots, p\}$

$$\begin{aligned}
\mathsf{E}(X_k^2) &= \frac{1}{p^2} \sum_{i,i'=1}^m \mathsf{E}\Big((1 - \frac{z_{ik}}{p_i})(1 - \frac{z_{i'k}}{p_{i'}})\Big) \psi_{i'} \psi_{i'}^\top \psi_i \psi_i^\top \\
&= \frac{1}{p^2} \sum_{i,i'=1}^m \Big(\frac{\mathsf{E}(z_{ik} z_{i'k})}{p_i p_{i'}} - 1\Big) \psi_{i'} \psi_{i'}^\top \psi_i \psi_i^\top \\
&= \frac{1}{p^2} \sum_{i=1}^m \Big(\frac{\mathsf{E}(z_{ik}^2)}{p_i^2} - 1\Big) \psi_i \psi_i^\top \psi_i \psi_i^\top \\
&= \frac{1}{p^2} \sum_{i=1}^m \Big(\frac{1}{p_i} - 1\Big) \|\psi_i\|_2^2 \psi_i \psi_i^\top \\
&\preceq \frac{1}{p^2} \sum_{i=1}^m \frac{\|\psi_i\|_2^2}{p_i} \psi_i \psi_i^\top.
\end{aligned}$$

Given that the probability distribution (p_i) verifies $p_i \geq \beta \frac{\|\psi_i\|_2^2}{\|\Psi\|_F^2}$, we get $\mathsf{E}(Y^2) \preceq \frac{\|\Psi\|_F^2}{\beta p} \sum_{i=1}^m \psi_i \psi_i^\top = \frac{\|\Psi\|_F^2}{\beta p} \Psi\Psi^\top$. Hence $\|\mathsf{E}(Y^2)\|_2 \leq \frac{\|\Psi\|_F^2}{\beta p} \lambda_{\max}(\Psi\Psi^\top)$. We now apply Theorem 5 with $R = \frac{1}{p} \lambda_{\max}(\Psi\Psi^\top)$ and $\sigma^2 = \frac{\|\Psi\|_F^2}{\beta p} \lambda_{\max}(\Psi\Psi^\top)$, which leads to the desired result.

C Proof of Theorem 3

Using Theorem 1, the fact that $L \preceq K$ and that the variance of the estimator $\hat{f}_K$ is matrix-increasing as a function of K (Eqn. (4)), it follows that

$$\begin{aligned}
\mathsf{E}_\xi \|\hat{f}_L - f^*\|_2^2 &= \text{bias}(L)^2 + \text{variance}(L) \\
&\leq \left(1 - \frac{\gamma/\lambda}{1-t}\right)^{-2} \text{bias}(K)^2 + \text{variance}(K) \\
&\leq \left(1 - \frac{\gamma/\lambda}{1-t}\right)^{-2} (\text{bias}(K)^2 + \text{variance}(K)) \\
&= \left(1 - \frac{\gamma/\lambda}{1-t}\right)^{-2} \mathsf{E}_\xi \|\hat{f}_L - f^*\|_2^2.
\end{aligned}$$

We set $\gamma = \lambda\epsilon$ and $t = 1/2$ (which verifies $\gamma \leq (1-t)\lambda$). The above holds if

$$\lambda_{\max}\Big(\Phi - \Phi^{1/2} U^\top S S^\top U \Phi^{1/2}\Big) \leq t.$$

Now let $\Psi = \Phi^{1/2} U^\top$. Then we have $\|\psi_i\|_2^2 = l_i(\gamma)$ and $\|\Psi\|_F^2 = d_{\text{eff}}$. Using Theorem 2 on Ψ, and given that $\lambda_{\max}(\Psi\Psi^\top) = \lambda_{\max}(\Phi) \leq 1$, for the result to hold with probability at least $1-\rho$, it is sufficient to set p such that $n \exp\left(\frac{-p(1/2)^2/2}{d_{\text{eff}}/\beta + 1/6}\right) \leq \rho$, which gives the desired lower bound $p \geq 8(d_{\text{eff}}/\beta + 1/6) \log\left(\frac{n}{\rho}\right)$.

D Proof of Theorem 4

First, it is clear that

$$\begin{aligned}
\tilde{l}_i &= e_i^\top B (B^\top B + n\lambda I)^{-1} B^\top e_i \\
&= e_i^\top B B^\top (B B^\top + n\lambda I)^{-1} e_i \\
&= \text{diag}(L(L + n\lambda I)^{-1})_i,
\end{aligned}$$

with e_i the i-th element of the standard basis in $\mathbb{R}^n$. Now we bound the approximations $\tilde{l}_i$ by comparing the matrices $L(L+n\lambda I)^{-1}$ and $K(K+n\lambda I)^{-1}$ with respect to the semidefinite order. Since $L \preceq K$ (see Eqn. (11) in Appendix A) and the map $K \to K(K+n\lambda I)^{-1}$ is matrix-increasing, we immediately get the upper bound $\tilde{l}_i \leq l_i(\lambda)$ for all $i \in \{1, \cdots, n\}$. Next, we derive the lower bound. For $\gamma > 0$, we consider again the regularized approximation

$$L_\gamma = KS(S^\top K S + n\gamma I)^{-1} S^\top K,$$

with $S \in \mathbb{R}^{n \times p}$ the sketching matrix. Due to the matrix inversion lemma, $L_\gamma \preceq L$ (see Eqn. (11) in Appendix A). Hence to get a lower bound on $\tilde{l}_i$, it suffices to obtain a lower bound for the same quantity when L is replaced by L_γ. We prove in Appendix A that if

$$\lambda_{\max}\Big(\Psi\Psi^\top - \Psi S S^\top \Psi^\top\Big) \leq t$$

for $t \geq 0$ with $\Psi = \Phi^{1/2} U^\top$, $\Phi = \Sigma(\Sigma + n\gamma I)^{-1}$, then

$$K - L_\gamma \preceq \frac{n\gamma}{1-t} K(K + n\gamma)^{-1} \preceq \frac{\gamma/\lambda}{1-t} I.$$

Therefore

$$L_\gamma(L_\gamma + n\lambda I)^{-1} \succeq (K - \frac{n\gamma}{1-t} I)(K + n\lambda I)^{-1}$$
$$\succeq K(K + n\lambda I)^{-1} - \frac{\gamma/\lambda}{1-t} I.$$

Hence $\tilde{l}_i \geq l_i(\lambda) - \frac{\gamma/\lambda}{1-t}$. Now we choose again $t = 1/2$ and $\gamma = \epsilon\lambda$ for $\epsilon \in (0, 1/2)$, we get the additive error bound on $\tilde{l}_i$ and we finish the proof as in Theorem 3 to get the lower bound $p \geq 8(d_{\text{eff}}/\beta + 1/6) \log\left(\frac{n}{\rho}\right)$, with $d_{\text{eff}}/\beta = \mathsf{Tr}(K)/(n\gamma)$. The latter equality follows from $\beta = n\lambda\epsilon d_{\text{eff}}/\mathsf{Tr}(K)$ (fourth remark after Theorem 4).
As for the multiplicative error bound, using $K - L_\gamma \preceq \frac{n\gamma}{1-t} K(K + n\gamma)^{-1}$ we get

$$L_\gamma(L_\gamma + n\lambda I)^{-1} \succeq (K - \frac{n\gamma}{1-t} K(K + n\gamma)^{-1})(K + n\lambda I)^{-1}$$
$$= K(K + n\lambda I)^{-1}(I - \frac{n\gamma}{1-t}(K + n\gamma I)^{-1}).$$

For $t = 1/2$, $I - \frac{n\gamma}{1-t}(K + n\gamma I)^{-1} = (K - n\gamma I)(K + n\gamma I)^{-1} \succeq \frac{\sigma_n - n\gamma}{\sigma_n + n\gamma} I$. The result follows.