Efficient Estimation of Kernel Matrix Spectral Norm using Random Features

Yiting Cao
School of Computer Science
University of Oklahoma
Norman, USA
yitingcao06@gmail.com

Shayan Shafaei
School of Computer Science
University of Oklahoma
Norman, USA
shayan.shafaei@ou.edu

Luyuan Yang
School of Computer Science
University of Oklahoma
Norman, USA
luyuan.yang@ou.edu

Chao Lan
School of Computer Science
University of Oklahoma
Norman, USA
clan@ou.edu

***Abstract*—This paper proposes a new approach to accelerate spectral norm estimation for a kernel matrix of n data points. Our key intuition is that, by applying the seminal random feature technique, we can well estimate the norm without computing or operating on the n-by-n kernel matrix but only an n-by-q random feature matrix with $q \ll n$ features, thereby significantly reducing the estimation time from $O(n^2)$ to $O(nq)$.**

Technically, our analysis suggests the spectral norm of a kernel matrix can be approximated by that of its corresponding random feature matrix with an $\tilde{O}(\ln n/\sqrt{q})$ relative norm approximation error. This is comparable to the relative norm estimation error of power iteration (PI), a popular efficient norm estimation method, and suggests our method can be integrated with PI to further accelerate norm estimation without deteriorating the estimation accuracy. Based on these insights, we design a random feature-based power iteration (RFPI) estimator for the kernel matrix spectral norm. Experimental results on two real-world datat sets show RFPI has significantly less estimation time than PI while maintaining competitive estimation accuracy.

***Index Terms*—random feature, kernel matrix, spectral norm**

I. Introduction

Kernel matrix is a main ingredient of the powerful kernel methods [1], [2], and its spectral norm plays an important role in analyzing the performance of many kernel methods such as multiple kernel learning [3], matrix-valued kernel learning [4] and kernel PCA [5]; there are also interests in analyzing the non-asymptotic [6], [7] and asymptotic [8] norm properties.

This paper considers a practical problem on kernel matrix spectral norm i.e., given n data points $x_1, \ldots, x_n$ sampled from a population X and a kernel function $k : X \times X \to \mathbb{R}$, how to efficiently estimate the spectral norm of their kernel matrix $K \in \mathbb{R}^{n\times n}$ where $K(i, j) = k(x_i, x_j)$?

Since spectral norm coincides with the top singular value, a basic approach is to apply singular value decomposition (SVD) on K to retrieve its singular values which normally consumes $O(n^3)$ time. A more popular and efficient alternative is to apply power iteration on K to only estimate its top singular value which normally consumes $O(n^2)$ time. In either case, an $O(n^2)$ estimation time seems inevitable since one needs to compute and operate on an n-by-n kernel matrix.

This paper reduces the estimation time to $O(nq)$ by applying the random feature technique [9], where q is the number of random features used to approximate the kernel function and is often much smaller than n. Our key intuition is that norm estimation can be performed without computing or operating on the kernel matrix but a much smaller n-by-q random feature matrix. To our knowledge, random feature has been widely studied for accelerating kernel machines e.g., [10], [11], [12], [13], [14], [15] but not for kernel matrix norm estimation.

TABLE I
Notations for Matrix M

$M(i, j)$	entry of M at row i column j
$M_{i:}$	row i of M
$M_{:j}$	column j of M
$\|\|M\|\|$	spectral norm of M
$\sigma_i(M)$	i_{th} singular value of M s.t. $\sigma_1(M) \geq \sigma_2(M) \geq \ldots$
$\lambda_i(M)$	i_{th} eigenvalue of M s.t. $\lambda_1(M) \geq \lambda_2(M) \geq \ldots$

This paper makes several technical contributions. First, we show the spectral norm of a kernel matrix can be approximated by that of a smaller random feature matrix with an $\tilde{O}(\frac{n}{\sqrt{q}})$ error using the matrix Bernstein's inequality. Second, we show this error implies an $\tilde{O}(\frac{\ln n}{\sqrt{q}})$ relative norm approximation error that is comparable to the relative norm estimation error of power iteration [16]. This suggests our method can be integrated with power iteration to further accelerate estimation without lowering estimation accuracy. Finally, based on the above insights, we design a novel random feature-based power iteration (RFPI) method to estimate kernel matrix spectral norm. Our experimental results on real-world data sets show RFPI is significantly more efficient than power iteration while maintaining competitive estimation accuracy.

II. Kernel Matrix Spectral Norm Approximation based on Random Features

We propose to approximate the spectral norm of a kernel matrix by that of a random feature matrix. In this section we analyze the approximation errors. Table I lists the main matrix notations used in the paper. Below are more preliminaries.

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be n data points and $K \in \mathbb{R}^{n \times n}$ be their kernel matrix associated with a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that $K(i,j) = k(x_i, x_j)$. Following [9], construct a random feature matrix $Z \in \mathbb{R}^{n \times q}$ such that

$$Z_{i:} = \sqrt{\frac{2}{q}} \big[\cos(w_1^T x_i + b_1), \ldots, \cos(w_q^T x_i + b_q) \big], \quad (1)$$

where $w_1, \ldots, w_q \in \mathbb{R}^p$ are sampled i.i.d. from some distribution p (determined by the kernel) and $b_1, \ldots, b_q$ are sampled i.i.d. in $[0, 2\pi]$. Our analysis will use the following tools.

Lemma II.1 (Weyl's Inequality)**.** *For any symmetric matrices $S, T \in \mathbb{R}^{n \times n}$, we have*

$$\max_{i \in \{1,\ldots,n\}} | \lambda_i(S) - \lambda_i(T) | \leq ||S - T||. \quad (2)$$

Lemma II.2 (Bernstein's Inequality)**.** *Let $E_1, \ldots, E_q \in \mathbb{R}^{n \times n}$ be independent and zero-mean random matrices such that $||E_i|| \leq c$ almost surely for each i. Then, for every $\varepsilon \geq 0$,*

$$\Pr\left\{ \left| \lambda_1 \left(\sum_{i=1}^q E_i \right) \right| \geq \varepsilon \right\} \leq 2n \exp\left(-\frac{\varepsilon^2/2}{\sigma^2 + c\varepsilon/3} \right), \quad (3)$$

where $\sigma^2 = || \sum_{i=1}^q \mathbb{E} E_i^2 ||$.

A. Main Result

Our main result states that $||K||$ can be well approximated $||Z||^2$, as specified in the following theorem.

Theorem II.3. *Suppose k is a bounded shift-invariant kernel. For any $\varepsilon > 0$, we have $| \, ||K|| - ||Z||^2 \, | \leq \varepsilon$ with probability at least $1 - 2n \exp(-\frac{\varepsilon^2 q}{c_1 n^2 + c_2 n \varepsilon})$ over the random sampling of Z, where $c_1, c_1 > 0$ are constants.*

Proof. First note $||Z||^2 = ||ZZ^T||$. Then Lemma II.1 implies

$$| \, ||K|| - ||ZZ^T|| \, | = |\lambda_1(K) - \lambda_1(ZZ^T)| \leq ||K - ZZ^T||. \quad (4)$$

We now aim to bound $||K - ZZ^T||$. Write

$$K - ZZ^T = \sum_{i=1}^q (\frac{1}{q} K - Z_{:i} Z_{:i}^T) = \sum_{i=1}^q E_i, \quad (5)$$

where $E_i = \frac{1}{q} K - Z_{:i} Z_{:i}^T$. We can show E_i's satisfy two conditions. First, each E_i is a zero-mean matrix because

$$E_i(a,b) = \frac{K(a,b) - 2\cos(w_i^T x_a + b_i)\cos(w_i^T x_b + b_i)}{q}, \quad (6)$$

and by design (see detailed arguments in [9])

$$\mathbb{E}\big[\sqrt{2}\cos(w_i^T x_a + b_i) \sqrt{2}\cos(w_i^T x_b + b_i) \big] = K(a,b). \quad (7)$$

Second, E_i and E_j are independent whenever $i \neq j$ because (w_i, b_i) and (w_j, b_j) are independently sampled. Based on both conditions, we can apply Lemma II.2 and have

$$\Pr\left\{ \left| \lambda_1 \left(\sum_{i=1}^q E_i \right) \right| \geq \varepsilon \right\} \leq 2n \exp(-\frac{\varepsilon^2/2}{\sigma^2 + c\varepsilon/3}) \quad (8)$$

where c is an upper bound of $||E_i||$ and $\sigma^2 = || \sum_{i=1}^q \mathbb{E} E_i^2 ||$.

Now we specify c and σ^2. For c, since (6) implies

$$||E_i|| \leq n \max_{a,b} |E_i(a,b)| \leq \frac{n(c_* + 2)}{q}, \quad (9)$$

where c_* is the bound of kernel, we can set $c = \frac{n(c_*+2)}{q}$.

For σ^2, we have

$$\sigma^2 = q||\mathbb{E} E_i^2|| \leq qn \max_{a,b} |\mathbb{E} E_i^2(a,b)| \leq \frac{n^2(c_*^2 + 4)}{q}, \quad (10)$$

where the last inequality is based on (7) which implies

$$\mathbb{E} E_i^2 = \mathbb{E}(Z_{:i} Z_{:i}^T)^2 - \frac{K^2}{q^2}, \quad (11)$$

and thus

$$\begin{aligned} & | \, \mathbb{E} E_i^2(a,b) \, | \\ &= \left| \sum_{j=1}^n \mathbb{E} Z(a,i) Z(b,i) Z(j,i)^2 - \frac{1}{q^2} K(a,j) K(j,b) \right| \\ &\leq \sum_{j=1}^n \left| \mathbb{E} Z(a,i) Z(b,i) Z(j,i)^2 - \frac{1}{q^2} K(a,j) K(j,b) \right| \\ &\leq n \max_j \left| \mathbb{E} Z(a,i) Z(b,i) Z(j,i)^2 - \frac{1}{q^2} K(a,j) K(j,b) \right| \\ &\leq \frac{n(c_*^2 + 4)}{q^2}, \end{aligned} \quad (12)$$

where the last inequality holds as $Z(a,b) \in [-\sqrt{\frac{2}{q}}, \sqrt{\frac{2}{q}}]$.

Plugging (9) and (12) back to (8) gives

$$\Pr\left\{ \left| \lambda_1 \left(\sum_{i=1}^q E_i \right) \right| \geq \varepsilon \right\} \leq 2n \exp(-\frac{\varepsilon^2 q}{c_1 n^2 + c_2 n \varepsilon}), \quad (13)$$

where $c_1 = 2(c_*^2 + 4)$ and $c_2 = 2(c_* + 2)/3$. Combining this with (4) and (5) proves the theorem. □

B. Implications of the Main Result

A practical implication of Theorem II.3 is we can estimate $||Z||^2$ as an approximation of $||K||$, and the former estimation is more efficient since Z is often much smaller than K.

On the theory side, note Theorem II.3 implies the following *relative* norm approximation error.

Corollary II.4. *If $||K|| = \Theta(n)$, then Theorem II.3 implies*

$$\Pr\left\{ \frac{\big| \, ||K|| - ||Z||^2 \, \big|}{||K||} > \varepsilon \right\} \leq \tilde{O}(n \exp(-\varepsilon^2 q)). \quad (14)$$

Proof. In Theorem II.3, replace ε with $\varepsilon ||K||$ gives a failure probability $\tilde{O}(n \exp(-\frac{\varepsilon^2 q ||K||^2}{n^2}))$. Plugging $||K|| = \Theta(n)$ into this probability proves the corollary. □

We can compare this error with the relative estimation error of power iteration in [16, Theorem 4.1], rephrased as follows.

Theorem II.5. *For any symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and $T \geq 2$, the power method with T updates gives an estimate of $||A||$, denoted as $||\tilde{A}||$, with*

$$\Pr\left\{\frac{\left|\|A\| - \|\tilde{A}\|\right|}{\|A\|} > \varepsilon\right\} \leq O(\sqrt{\frac{(1-\varepsilon)^{2T}}{\varepsilon T}}). \tag{15}$$

Now we compare the two guarantees. Corollary II.4 suggests our norm approximation method has an $\tilde{O}(\frac{\ln n}{\sqrt{q}})$ relative error. In Theorem II.5, by relaxing $(1-\varepsilon)^T \leq \exp(-\varepsilon T)$, we see the power iteration norm estimation method has an $\tilde{O}(\frac{\ln n}{T})$ relative error. Note that both errors have the same logarithmic dependence on n, suggesting our method can be integrated with power iteration to further accelerate norm estimation without significantly lowering its estimation accuracy (as n increases). This motivates us to design a random feature-based power iteration method to estimate the kernel matrix spectral norm, which is presented in the next section.

III. Random Feature-based Power Iteration

Our designed random feature-based power iteration (RFPI) estimation method is presented in Algorithm 1. Note it consumes $O(nq)$ time as each of Steps 1, 3 and 4 does so, which is more efficient than the $O(n^2)$ time of power iteration. It also inherits the following standard convergence guarantee.

Lemma III.1. *In RFPI, $||\tilde{Z}||$ converges to $||Z||$ as $T \to \infty$.*

Proof. Let $v^{(t)}$ be v after t rounds of update. We will show it converges to the right singular vector of Z associated with the top singular value. Let $v^{(0)}$ be the randomly initialized v. By the update rules we have

$$v^{(t)} = \frac{(Z^TZ)^t \cdot v^{(0)}}{\|(Z^TZ)^t \cdot v^{(0)}\|_F}. \tag{16}$$

Let $v_1, \cdots, v_q$ be a set of orthonormal eigenvectors of Z^TZ and $\lambda_1, \cdots, \lambda_q$ be the associated eigenvalues such that $\lambda_1 > \lambda_2 > \ldots$. Since v_i's form a basis of $\mathbb{R}^q$, there exist some constants $c_1, \cdots, c_q \in \mathbb{R}$ such that

$$v^{(0)} = c_1v_1 + \cdots + c_qv_q. \tag{17}$$

This implies

$$\left(Z^TZ\right)^t v^{(0)} = \sum_{i=1}^{q} c_i \left(Z^TZ\right)^t v_i = \sum_{i=1}^{q} c_i\lambda_i^t v_i. \tag{18}$$

Plugging this back to (16) gives

$$v^{(t)} = \frac{c_1v_1 + c_2\frac{\lambda_2^t}{\lambda_1^t}v_2 + \cdots + c_q\frac{\lambda_q^t}{\lambda_1^t}v_q}{||c_1v_1 + c_2\frac{\lambda_2^t}{\lambda_1^t}v_2 + \cdots + c_q\frac{\lambda_q^t}{\lambda_1^t}v_q||_F}. \tag{19}$$

It is clear that (19) converges to v_1 as $t \to \infty$ if λ_1 is unique. In fact, with minor modifications of the arguments, we can show the conclusion holds when λ_1 is not unique.

Let $u^{(t)}$ be u after t rounds of update. By similar arguments, we can show it converges to the top left singular vector of Z. This completes the proof. □

Algorithm 1 The RFPI Algorithm

Input: data points $x_1, \ldots, x_n \in \mathbb{R}^p$, hyper-parameter q

1: Compute random feature matrix $Z \in \mathbb{R}^{n\times q}$ based on (1)

2: Randomly initialize $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^q$

for iteration = 1, …, T **do**

3: Update $v = \frac{Z^Tu}{||Z^Tu||_F}$

4: Update $u = \frac{Zv}{||Zv||_F}$

end for

Output: $||\tilde{Z}|| := (u^TZv)^2$ as an estimate of $||Z||$.

Note in Algorithm 1, applying power iteration to estimate the top singular value of the rectangular matrix Z is crucial. Although in theory it is equivalent (and perhaps more popular) to apply power iteration to estimate the top eigenvector of ZZ^T, in practice the former is faster as it avoids computing ZZ^T – again, this reinforces the key intuition of our proposed accelerated estimator, which is to avoid computing the kernel matrix and rely on the random feature matrix solely.

IV. Experiment

We experiment on the Communities and Crime data set and White Wine Quality data set. On each set, we treat the first n instances as input data and estimate the spectral norm of their kernel matrix. The following methods are evaluated.

- SVD: apply SVD on the kernel matrix. The result will be treated as ground truth to evaluate estimation accuracy.
- PI: apply power iteration on the kernel matrix.
- RFPI: the proposed method in Algorithm 1.

We use the Gaussian kernel $k(x_i, x_j) = \exp(-\frac{||x_i - x_j||_F^2}{2\sigma^2})$ with $\sigma = 0.1$. For RFPI, w is sampled from $N(0, I)$ where I is an identity matrix. All features are standardized and all reported results are averaged over 50 random trials.

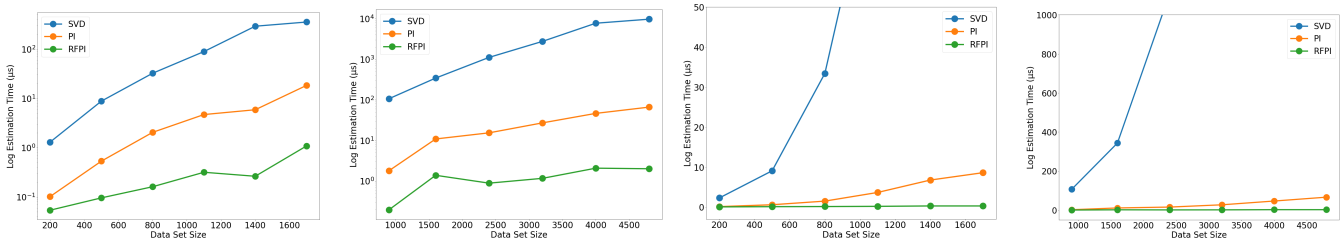
A. Estimation Performance versus Data Size

We first evaluate the impact of input data size on estimation. For RFPI, we fix q to 50 on Crime and 150 on Wine, and the number of power updates to 5. These hyper-parameters are chosen based on our sensitivity analysis in the next section.

Figure 1 shows the impact on estimation time. We see RFPI is more efficient than PI and SVD, and the gap increases as data size increases. Moreover, RFPI time scales almost linearly as data size increases whereas, comparatively, PI time scales quadratically and SVD time scales cubically. These coincide with our computational complexity analysis.

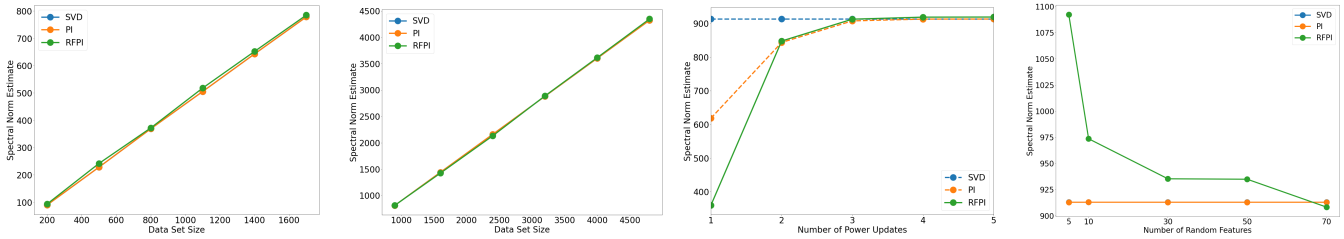
Figures 2(a) and 2(b) show the estimation accuracy. We see RFPI is very close to the PI and SVD across different data sizes, which verifies its effectiveness and coincides with our theoretical implication that RFPI and PI have close error rates.

Detailed estimates on both data sets are listed in Tables II and III. We see RFPI is close to both SVD and PI on average, but we also observe that its estimate has a large variance e.g., often 5∼10% of the estimate. We believe this variance is partly



(a) Time on Crime (log-scale) (b) Time on Wine (log-scale) (c) Time on Crime (linear-scale) (d) Time on Wine (linear-scale)

Fig. 1. Estimation Time versus Data Size



(a) Norm Estimate on Crime (b) Norm Estimate on Wine (c) Estimate vs Power Updates (d) Estimate vs # Random Features

Fig. 2. Estimation Performance

TABLE II
NORM ESTIMATES ON CRIME

n	**.2k**	**.5k**	**.8k**	**1.1k**	**1.4k**	**1.7k**
SVD	90	228	369	505	643	778
PI	90	228	369	505	643	778
RFPI	92	229	390	507	644	784

TABLE III
NORM ESTIMATES ON WINE

n	**.9k**	**1.6k**	**2.4k**	**3.2k**	**4.0k**	**4.8k**
SVD	.81k	1.44k	2.16k	2.88k	3.60k	4.32k
PI	.81k	1.44k	2.16k	2.88k	3.60k	4.32k
RFPI	.83k	1.43k	2.17k	2.88k	3.58k	4.38k

inherited from the random feature technique, since it decreases as the number of random features increases, establishing a somewhat reasonable tradeoff between estimation variance and efficiency. Still, how to reduce this tradeoff and estimation variance remains an open challenge for RFPI.

B. Sensitivity Analysis

Figure 2(c) shows the estimate versus the number of power updates on Crime. We see both PI and RFPI converge after 5 updates, which justifies our choice in previous experiments.

Figure 2(d) shows the estimate versus the number of random features q on Crime. We treat all data points as input and report results averaged over 100 random trials. We see RFPI becomes close to the baseline at $q = 50$ on Crime, which is way smaller than the input data size $n = 1993$. More interestingly, we see RFPI converges at a rate close to $\tilde{Q}(1/\sqrt{q})$, which coincides with the implication of our theoretical guarantee.

V. CONCLUSION AND FUTURE WORK

This paper presents a novel approach to efficiently estimate the spectral norm of a kernel matrix for n data points that consumes only $O(n)$ time. Its main idea is to approximate the spectral norm of an n-by-n kernel matrix using that of a much smaller n-by-q random feature matrix. Our analysis suggests an $\tilde{O}(\frac{n}{\sqrt{q}})$ absolute norm approximation error and an $\tilde{O}(\frac{\ln n}{\sqrt{q}})$ relative norm approximation error. The latter is comparable to the relative estimation error of power iteration (PI), suggesting our method can be integrated with PI to further accelerate estimation without sacrificing estimation accuracy. Motivated by this observation, we design a random feature-based power iteration (RFPI) estimator which consumes only $O(nq)$ time in contrast to the $O(n^2)$ time of power iteration. Our experimental results on two data sets show RFPI is significantly more efficient than power iteration while maintaining competitive estimation accuracy.

It is worth mentioning this paper explores a new direction to accelerate kernel matrix norm estimation, which aims to avoid computing the kernel matrix while existing solutions such as power iteration only speed up estimation *after* computing the kernel matrix. Therefore, the proposed idea may be integrated with other efficient norm estimators such as based on Nystrom method [17], [18] or leverage score sampling [19]. It may also be interesting to explore the use of data/model-dependent random feature mappings [20], [21], [22] especially those specifically designed to approximate kernels e.g., [23], [24].

REFERENCES

[1] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[2] J. Shawe-Taylor, "Kernel methods for pattern analysis," *Cambridge University Press google schola*, vol. 2, pp. 181–201, 2004.

[3] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine learning research*, vol. 5, no. Jan, pp. 27–72, 2004.

[4] V. Sindhwani, M. H. Quang, and A. C. Lozano, "Scalable matrix-valued kernel learning for high-dimensional nonlinear multivariate regression and granger causality," *arXiv preprint arXiv:1210.4792*, 2012.

[5] J. Shawe-Taylor, C. K. Williams, N. Cristianini, and J. Kandola, "On the eigenspectrum of the gram matrix and the generalization error of kernel-pca," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2510–2522, 2005.

[6] L. Jia and S. Liao, "Accurate probabilistic error bound for eigenvalues of kernel matrix," in *Asian Conference on Machine Learning*. Springer, 2009, pp. 162–175.

[7] S. P. Kasiviswanathan and M. Rudelson, "Spectral norm of random kernel matrices with applications to privacy," *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, p. 898, 2015.

[8] Z. Fan and A. Montanari, "The spectral norm of random inner-product kernel matrices," *Probability Theory and Related Fields*, vol. 173, pp. 27–85, 2019.

[9] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Advances in neural information processing systems*, vol. 20, 2007.

[10] F. X. X. Yu, A. T. Suresh, K. M. Choromanski, D. N. Holtmann-Rice, and S. Kumar, "Orthogonal random features," *Advances in neural information processing systems*, vol. 29, 2016.

[11] F. Bach, "On the equivalence between kernel quadrature rules and random feature expansions," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 714–751, 2017.

[12] T. Dao, C. M. De Sa, and C. Ré, "Gaussian quadrature for kernel features," *Advances in neural information processing systems*, vol. 30, 2017.

[13] M. Munkhoeva, Y. Kapushev, E. Burnaev, and I. Oseledets, "Quadrature-based features for kernel approximation," *Advances in neural information processing systems*, vol. 31, 2018.

[14] H. Yamasaki, S. Subramanian, S. Sonoda, and M. Koashi, "Learning with optimized random features: Exponential speedup by quantum machine learning without sparsity and low-rank assumptions," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 674–13 687, 2020.

[15] Y. Cao and C. Lan, "Active approximately metric-fair learning," in *Uncertainty in Artificial Intelligence*. PMLR, 2022, pp. 275–285.

[16] J. Kuczyński and H. Woźniakowski, "Estimating the largest eigenvalue by the power and lanczos algorithms with a random start," *SIAM journal on matrix analysis and applications*, vol. 13, no. 4, pp. 1094–1122, 1992.

[17] P. Drineas, M. W. Mahoney, and N. Cristianini, "On the nyström method for approximating a gram matrix for improved kernel-based learning." *journal of machine learning research*, vol. 6, no. 12, 2005.

[18] T. Yang, Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou, "Nyström method vs random fourier features: A theoretical and empirical comparison," *Advances in neural information processing systems*, vol. 25, 2012.

[19] M. B. Cohen, C. Musco, and C. Musco, "Input sparsity time low-rank approximation via ridge leverage score sampling," in *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2017, pp. 1758–1777.

[20] Y. Xu, H. Yang, L. Zhang, and T. Yang, "Efficient non-oblivious randomized reduction for risk minimization with improved excess risk guarantee," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[21] Y. Cao and C. Lan, "A model-agnostic randomized learning framework based on random hypothesis subspace sampling," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[22] J. Sturges, L. Yang, S. Shafaei, and C. Lan, "Efficient data-dependent random projection for least square regressions," *International Conference on Acoustics, Speech, and Signal Processing*, 2025.

[23] J. Yang, V. Sindhwani, H. Avron, and M. Mahoney, "Quasi-monte carlo feature maps for shift-invariant kernels," in *International Conference on Machine Learning*. PMLR, 2014, pp. 485–493.

[24] Z. Huang, J. Sun, and Y. Huang, "Quasi-monte carlo features for kernel approximation," in *Forty-first International Conference on Machine Learning*.