Approximating Gram Matrix Spectral Norm using Random Features with Applications in Efficient Norm Estimation

Anonymous authors
Paper under double-blind review

Abstract

This paper considers spectral norm estimation for the Gram matrix of n data points. Efficient estimation methods like power iteration or Nystrom directly work with the Gram matrix or submatrix and thus have their time complexities quadratically dependent on data size. This paper investigates an orthogonal direction to accelerate the estimation through norm approximation. Building on the seminal work of random features for kernel methods, we propose to approximate the spectral norm of Gram matrix by the spectral norm of a random feature matrix, which is often much smaller and hence more efficient to work with. Original theoretical analysis suggests the approximation has an $\tilde{O}(n/\sqrt{q})$ absolute error and $\tilde{O}(\ln n/\sqrt{q})$ relative error with q random features, close to the errors of prior methods. We apply the approximation to accelerate power iteration and Nystrom, improving their time complexities by replacing the quadratic dependence on data size with a linear dependence. Experimental results on two data sets show the accelerated methods significantly reduce the estimation time while being able to maintain the estimation accuracy.

1 Introduction

In machine learning, spectral norm is a fundamental tool widely used to design learning algorithms (Achlioptas & McSherry, 2005; Mazumder et al., 2010; Shivanna et al., 2015; Bietti et al., 2019; Roth et al., 2020; Zhang et al., 2021; Chen et al., 2021; Do & Luong, 2021) and characterize learning performance (Drineas et al., 2005; Kumar & Kannan, 2010; Takác et al., 2013; Gittens & Mahoney, 2013; Bartlett et al., 2017; Xiao et al., 2020). An efficient estimation of spectral norm is a cornerstone of efficient learning and analysis.

Another fundamental tool in machine learning is Gram matrix, whose efficient approximation (Rahimi & Recht, 2007; Drineas et al., 2005; Holodnak & Ipsen, 2015) and spectrum analysis (Shawe-Taylor et al., 2002; Hachem et al., 2005; Hoyle & Rattray, 2004; Benaych-Georges & Couillet, 2016) have been widely studied and used to characterize the generalization performance of kernel machines (Schölkopf et al., 1999; Shawe-Taylor et al., 2005; Suzuki, 2018; Ma et al., 2020).

This paper considers spectral norm estimation for the Gram matrix of n data points. A basic approach is to apply singular value decomposition on the Gram matrix to retrieve its eigenvalues and pick the largest one (equal to spectral norm), which consumes $O(n^3)$ time. A faster alternative is to apply power iteration (Epperson, 2021) on the Gram matrix to estimate its top eigenvalue, which consumes $O(n^2)$ time. Another fast alternative is based on Nystrom method (Drineas et al., 2005; Yang et al., 2012), which randomly samples r ($r < n$) data points from the input and estimates the scaled spectral norm of their Gram matrix to approximate the target norm – the estimation takes $O(r^3)$ time if done by SVD and $O(r^2)$ by power iteration.

Both power and Nystrom methods are very efficient and popular, yet they directly work with Gram matrices. In the seminal work (Rahimi & Recht, 2007), it is shown that a kernel function can be well approximated by the product of two random mappings. This motivates us to hypothesize the spectral norm of a Gram matrix (associated with the kernel) can also be well approximated by the spectral norm of a random mapping matrix. Most importantly, the mapping matrix is often way smaller than the Gram matrix, allowing its norm estimation to be performed more efficiently and thus further accelerating the power and Nystrom methods. To the best of our knowledge, although random feature has been widely studied e.g., (Yu et al., 2016; Bach,

2017; Dao et al., 2017; Munkhoeva et al., 2018; Yamasaki et al., 2020), its connection to spectral norm estimation has not been investigated or exploited before.

This paper proposes to approximate Gram matrix spectral norm using random features (Rahimi & Recht, 2007) and apply it to accelerate the power and Nystrom based norm estimation methods. Original theoretical analysis suggests the norm approximation quality is high i.e., with q random features, the approximation endures an $O(n/\sqrt{q})$ absolute error that matches the error of kernel function approximation in terms of q (Rahimi & Recht, 2008; Yang et al., 2012) and an $\tilde{O}(\ln n/\sqrt{q})$ relative error that is very close to the estimation error of power iteration in terms of n (Komzsik, 2003). On the efficiency side, the accelerated power and Nystrom based estimation methods only consume $O(n)$ and $O(r)$ time, respectively, significantly improving over the current $O(n^2)$ and $O(r^2)$ time respectively. We empirically evaluate the accelerated norm estimation methods on two real-world data sets and results show they significantly reduce estimation time while maintaining estimation accuracy. In particular, we observe their computation time indeed scale linearly as the data size increases, while their basic counterparts' time scale quadratically.

2 Spectral Norm Approximation based on Random Fourier Features

2.1 Notations

For matrix M, let $M(i,j)$ be its element at row i column j, $M_{i:}$ be its i_{th} row vector and $M_{:j}$ be its j_{th} column vector; let $||M||$ be its operator norm, $\sigma_i(M)$ be its i_{th} singular value such that $\sigma_1(M) \geq \sigma_2(M) \geq \ldots$ and, whenever eigenvalues exist, $\lambda_i(M)$ be its i_{th} eigenvalue such that $\lambda_1(M) \geq \lambda_2(M) \geq \ldots$. Note $||M|| = \lambda_1(M)$.

2.2 Main Result

Consider the Gram matrix $K \in \mathbb{R}^{n\times n}$ of n data points $x_1, \ldots, x_n \in \mathbb{R}^p$ associated with a kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that $K(i,j) = k(x_i, x_j)$. Construct a random feature matrix $Z \in \mathbb{R}^{n\times q}$ such that

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

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 main result is stated as follows.

Theorem 2.1. *Suppose K is constructed based on a bounded shift-invariant kernel. Then, for any $\varepsilon > 0$, we have $|\,||K|| - \sigma_1^2(Z)\,| \leq \varepsilon$ 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 depending on the kernel bound.*

Proof. We will apply the following two inequalities to prove the theorem.

Lemma 2.2 (Weyl's Inequality)**.** *For any symmetric matrices $S, T \in \mathbb{R}^{n\times n}$,*

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

Lemma 2.3 (Matrix 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 $\varepsilon \geq 0$, we have*

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

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

To prove Theorem 2.1, first note $\sigma_1^2(Z) = ||ZZ^T||$ and by the Weyl's inequality,

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

Thus we can focus on bounding $||K - ZZ^T||$. Write

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

where $E_i = \frac{1}{q} K - Z_{:i} Z_{:i}^T$. Note each E_i is a zero-mean matrix because

$$E_i(a,b) = \frac{1}{q} [K(a,b) - 2\cos(w_i^T x_a + b_i)\cos(w_i^T x_b + b_i)], \tag{6}$$

and by the Bochner's theorem (see detailed arguments in (Rahimi & Recht, 2007))

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

Moreover, for $i \neq j$, E_i and E_j are independent since (w_i, b_i) and (w_j, b_j) are independently sampled. Based on these conditions, we can apply the Matrix Bernstein's inequality on $\sum_{i=1}^{q} E_i$ and have

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

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

The remaining task is to specify or bound c and σ^2.

We can set $c = \frac{n(c_*+2)}{q}$ where $c_* > 0$ is an upper bound of the kernel function since (6) implies

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

For σ^2, we have

$$\left|\left| \sum_{i=1}^{q} \mathbb{E} E_i^2 \right|\right| = q \left|\left| \mathbb{E} E_i^2 \right|\right| \leq qn \max_{a,b} |\mathbb{E} E_i^2(a,b)| \leq \frac{n^2(c_*^2+4)}{q}. \tag{10}$$

The last inequality in the above argument is based on (7) so that

$$\mathbb{E} E_i^2 = \mathbb{E}\left[\frac{1}{q^2} K^2 + (Z_{:i} Z_{:i}^T)^2 - \frac{2}{q} K Z_{:i} Z_{:i}^T \right] = \mathbb{E}(Z_{:i} Z_{:i}^T)^2 - \frac{K^2}{q^2}, \tag{11}$$

and thus

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

where the last inequality is due to the fact that $Z(a,b) \in [-\sqrt{2/q}, \sqrt{2/q}]$ by design.

Plugging (9) and (12) back to (8) and merging constants, we have

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

where $c_1 = 2(c_*^2 + 4)$ and $c_2 = 2(c_* + 2)/3$.

Combining the above with (4) proves the theorem. □

2.3 Implications of the Main Result

A direct implication of Theorem 2.1 is we can speed up the estimation of $||K||$ by estimating $\sigma_1(Z)$ instead and using it to approximate $||K||$. The efficiency gain comes from the fact that Z is often smaller than K and therefore (i) Z can be computed faster than K from the input data points $x_1, \ldots, x_n$ and (ii) $\sigma_1(Z)$ can be estimated faster than $||K||$ based on proper estimation methods.

On the other hand, fixing the failure probability in Theorem 2.1 suggests ZZ^T has an $\tilde{O}(n/\sqrt{q})$ spectral norm approximation error. We can discuss its implications from two perspectives.

First, since random feature was initially introduced to approximate kernel function instead of Gram matrix spectral norm, we can compare our norm approximation error with the following kernel approximation error derived in (Rahimi & Recht, 2008; Yang et al., 2012).

Lemma 2.4. *Let M be a compact subset of $\mathbb{R}^p$ with diameter $d(M)$ and k be a shift-invariant kernel. Then for any $x_i, x_j \in M$, we have*

$$\Pr[\sup_{x_i, x_j} |z(x_i)^T z(x_j) - k(x_i, x_j)| \geq \varepsilon] \leq 2^8 \left(\frac{\sigma_p d(M)}{\varepsilon}\right)^2 \exp\left(-\frac{q\varepsilon^2}{4(p+2)}\right), \tag{14}$$

where σ_p^2 is the second moment of the Fourier transform of k.

This lemma suggests an $\tilde{O}(1/\sqrt{q})$ kernel approximation error which has the same dependence on q as our norm approximation error, implying the latter error is largely inherited from the former error. We notice (Rudi & Rosasco, 2017) suggests the kernel approximation error could be further improved to $\tilde{O}(1/q)$. Whether such improvement also applies to norm approximation error remains an open problem.

Second, we can compare our approximation quality with that of the classic power method (which continuously updates a randomly initialized vector $z \in \mathbb{R}^n$ by $z = Kz$ and at last estimates the spectral norm by $z^T Kz$). The latter's relative norm approximation error has been well studied, and an example in (Kuczyński & Woźniakowski, 1992, Theorem 4.1) is stated as follows.

Theorem 2.5. *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 $||A||$ using $||\tilde{A}||$ with*

$$\Pr\left\{\frac{\left|\|A\| - \|\tilde{A}\|\right|}{\|A\|} > \varepsilon\right\} \leq \min(0.824, \frac{0.354}{\sqrt{\varepsilon(T-1)}})\sqrt{n}(1-\varepsilon)^{T-1/2}. \tag{15}$$

Accordingly, we can extend Theorem 2.1 to obtain a *relative* norm approximation error guarantee.

Corollary 2.6. *Suppose K is constructed with a bounded shift-invariant kernel and $||K|| = \Theta(n)$. Then*

$$\Pr\left\{\frac{\left|||K|| - ||ZZ^T|||\right|}{||K||} > \varepsilon\right\} \leq \tilde{O}(n \exp(-\varepsilon^2 q)). \tag{16}$$

Proof. In Theorem 2.1, replacing ε with $\varepsilon||K||$ gives a failure probability $\tilde{O}(n \exp(-\frac{\varepsilon^2 q||K||^2}{n^2}))$. Plugging in $||K|| = \Theta(n)$ proves the corollary. □

Now we compare the two guarantees. Corollary 2.6 suggests our method has an $\tilde{O}(\frac{\ln n}{\sqrt{q}})$ relative error rate. In Theorem 2.5, by relaxing $(1-\varepsilon)^T \leq \exp(-\varepsilon T)$, we see the power method has a roughly $\tilde{O}(\frac{\ln n}{T})$ relative error rate. Both rates have the same logarithm dependence on n, implying our method is as robust as power method when data size scales up.

Note, however, it is not exactly fair to compare our method with power method since they speed up norm estimation from orthogonal directions: we replace the target matrix K with a smaller Z, while power method replaces the traditional estimator (for any target matrix) with a faster one.

3 Applying the Approximation to Speed up Spectral Norm Estimation

In this section, we demonstrate how to Theorem 2.1 to speed up the estimation of $||K||$ given a set of data points $x_1, \ldots, x_n$. We focus on accelerating two popular estimation methods: power iteration and Nystrom method, and will refer to their accelerated methods as RFF-Power and RFF-Nystrom, respectively.

3.1 Power versus RFF-Power

A popular method to estimate the top eigenvalue of a matrix is power iteration. To apply it, we first compute the Gram matrix K from n data points in $O(n^2)$ time and then iteratively update a randomly initialized vector $z \in \mathbb{R}^n$ by $z = \frac{Kz}{||Kz||_F}$ in $O(n^2)$ time, where $||\cdot||_F$ denotes the Frobenius norm. After that, $||K||$ is estimated as $z^T Kz$ in $O(n^2)$ time. Overall, the computational time has a quadratic dependence on n.

Based on Theorem 2.1, we can estimate $\sigma_1(Z)$ instead. We can first compute Z from n data points in $O(nq)$ time and then iteratively update two randomly initialized vectors $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^q$ by

$$v = \frac{Z^T u}{||Z^T u||_F} \quad \text{and} \quad u = \frac{Zv}{||Zv||_F}, \tag{17}$$

in $O(nq)$ time. After that, $\sigma_1(Z)$ is estimated as $(u^T Zv)^2$ in $O(nq)$ time. Overall, the computational time of RFF-Power has a linear dependence on n, more scalable than the standard Power method as n increases.

Convergence of the updates in (17) is guaranteed by the following lemma.

Lemma 3.1. *If $||Z||$ is unique and positive, then $(u^T Zv)^2$ converges to $||Z||$ as updates (17) proceed.*

Proof. Let $v^{(t)}$ and $u^{(t)}$ respectively denote the updated v and u after t rounds of (17). We aim to show $v^{(t)}$ converges to the right singular vector of Z associated with the top singular value, and $u^{(t)}$ converges to the left singular vector of Z associated with the top singular value.

We first study $v^{(t)}$. Let $v^{(0)}$ denote the randomly initialized v. By the update rules we have

$$v^{(t)} = \frac{(Z^T Z)^t \cdot v^{(0)}}{\|(Z^T Z)^t \cdot v^{(0)}\|_F}. \tag{18}$$

Let $v_1, \cdots, v_q$ be a set of orthonormal eigenvectors of $Z^T Z \in \mathbb{R}^{q \times q}$ and $\lambda_1, \cdots, \lambda_q$ be the associated eigenvalues satisfying $\lambda_1 > \lambda_2 > \ldots$. 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

$$v^{(0)} = c_1 v_1 + \cdots + c_q v_q. \tag{19}$$

By left multiplying $(Z^T Z)^t$ on both sides and the fact $\lambda_1 > 0$, we have

$$\left(Z^T Z\right)^t v^{(0)} = \sum_{i=1}^{q} c_i \left(Z^T Z\right)^t v_i = \sum_{i=1}^{q} c_i \lambda_i^t v_i = \lambda_1^t \left(c_1 v_1 + c_2 \frac{\lambda_2^t}{\lambda_1^t} v_2 + \cdots + c_q \frac{\lambda_q^t}{\lambda_1^t} v_q \right). \tag{20}$$

This implies

$$v^{(t)} = \frac{\left(Z^\top Z\right)^t v_0}{\| \left(Z^\top Z\right)^t v_0 \|_F} = \frac{\lambda_1^t \left(c_1 v_1 + c_2 \frac{\lambda_2^t}{\lambda_1^t} v_2 + \cdots + c_q \frac{\lambda_q^t}{\lambda_1^t} v_q \right)}{||\lambda_1^t \left(c_1 v_1 + c_2 \frac{\lambda_2^t}{\lambda_1^t} v_2 + \cdots + c_q \frac{\lambda_q^t}{\lambda_1^t} v_q \right)||_F} = \frac{c_1 v_1 + c_2 \frac{\lambda_2^t}{\lambda_1^t} v_2 + \cdots + c_q \frac{\lambda_q^t}{\lambda_1^t} v_q}{||c_1 v_1 + c_2 \frac{\lambda_2^t}{\lambda_1^t} v_2 + \cdots + c_q \frac{\lambda_q^t}{\lambda_1^t} v_q||_F}. \tag{21}$$

As t increases, the right-hand-side converges to $\frac{c_1 v_1}{||c_1 v_1||_F} = v_1$ because $\lambda_i < \lambda_1$ for all $i > 1$. This proves $v^{(t)}$ converges to the top eigenvector of $Z^T Z$ and thus the top right singular vector of Z.

By similar argument, $u^{(t)}$ converges to the top left singular vector of Z. The proof is complete. □

3.2 Nystrom versus RFF-based Nystrom

The Nystrom method offers a powerful way to construct the low-rank approximation of a matrix based on its randomly sampled submatrices (Drineas et al., 2005; Yang et al., 2012). Its idea can also be applied to estimate $||K||$ based on n input data points as follows: first, randomly sample r points from the input and construct their Gram matrix $\tilde{K} \in \mathbb{R}^{r \times r}$; then, estimate $\frac{n}{r}||\tilde{K}||$ as an approximation of $||K||$. Overall, the Nystrom process consumes $O(r^2)$ time for computing $\tilde{K}$ and $O(r^2)$ time for estimating $||\tilde{K}||$ using the power method. The total time has a quadratic dependence on r.

Based on Theorem 2.1, we can estimate $\sigma_1(\tilde{Z})$ instead of $||\tilde{K}||$, where $\tilde{Z} \in \mathbb{R}^{r \times q}$ is the random feature matrix whose rows are constructed based on (1) for the r sampled data points. Overall, the RFF-Nystrom process consumes $O(rq)$ time for computing $\tilde{Z}$ and $O(rq)$ time for estimating $\sigma_1(\tilde{Z})$ using the power method. The total time has a linear dependence on r, which is more scalable than Nystrom as r increases.

In the following, we will refer to $\frac{r}{n}$ as the data sampling ratio of Nystrom-based estimation methods.

4 Experiment

We experiment on two public real-world data sets, namely, the Communities and Crime data set and the White Wine Quality data set. On each data set, we pick the first n instances (only features, not labels), treat them as the input data and estimate the spectral norm of their Gram matrix using the following methods.

– SVD (baseline): compute the Gram matrix of, apply SVD on it and retrieve the top eigenvalue.

– Power: the standard power iteration method described in Section 3.1.

– RFF-Power: the accelerated power iteration method described in Section 3.1.

– Nystrom: the Nystrom method described in Section 3.2.

– RFF-Nystrom: the accelerated Nystrom method described in Section 3.2.

For all methods, we use the common Gaussian kernel function $k(x_i, x_j) = \exp(-\frac{||x_i - x_j||_F^2}{2\sigma^2})$ and set σ to 0.5 on Crime and 0.1 on Wine. For RFF-based methods, the sampling distribution (associated with Gaussian kernel) for w is $N(0, I)$ where I is an identity matrix. For Nystrom based methods, we pick the first r instances (out of n) to form the downsampled data set. All features are standardized to enhance numerical stability and all reported results are averaged over 50 random trials (except the last two which are over 100 random trials).

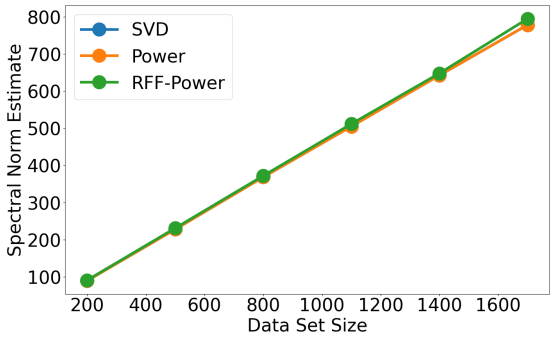
4.1 Spectral Norm Estimation versus Input Data Size

Performance of all methods versus input data size n are shown in Figure 1, where estimation time includes the time to compute Gram matrix, its submatrix or random feature matrix. For RFF-Power and RFF-Nystrom, we set the number of random features to 50 on Crime and 150 on Wine. For Power and RFF-Power, we set the number of power updates to 5. For Nystrom and RFF-Nystrom, we set the data sampling ratio to 0.8. These choices are based on sensitivity analysis results in Figure 2.

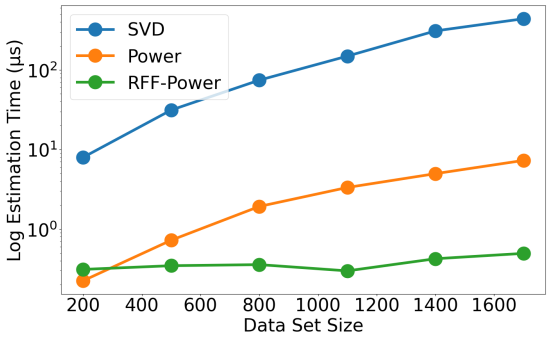
In Figure 1 (a)(c), we see the norm estimates of both Power and RFF-Power are very close to the SVD estimate, demonstrating their effectiveness. Figure 1 (b)(d) show how the log estimation time scales as n increases. We see RFF-Power consistently consumes the least amount of time, demonstrating its efficacy. Interesting, the time of three methods have roughly cubic, quadratic and linear dependence on data size, consistent with our theoretical analysis on their time complexities.

In Figure 1 (e)(g), we see the norm estimates of Nystrom and RFF-Nystrom are very close but slightly lower than the SVD estimate. We conjecture this gap is induced from the subsampling process of Nystrom and not from the application of RFF-based norm approximation. Figure 1 (f)(h) show how the estimation time scales as n increases. Again, we see RFF-Nystrom consistently consumes the least amount of time.

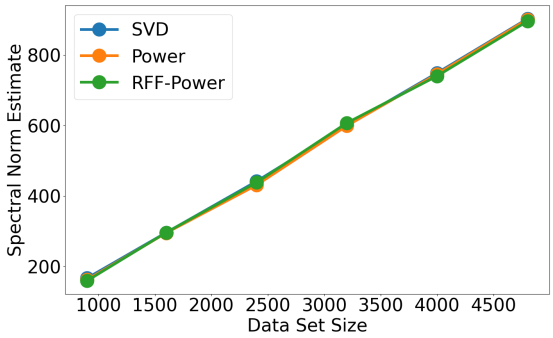
Overall, by comparing the left columns of Figure 1 and its right columns, we see RFF-Power and RFF-Nystrom speed up Power and Nystrom, respectively, without sacrificing estimation accuracy.



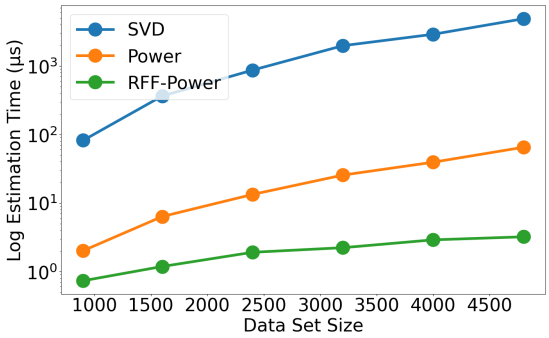
(a) Power-based Norm Estimates on Crime



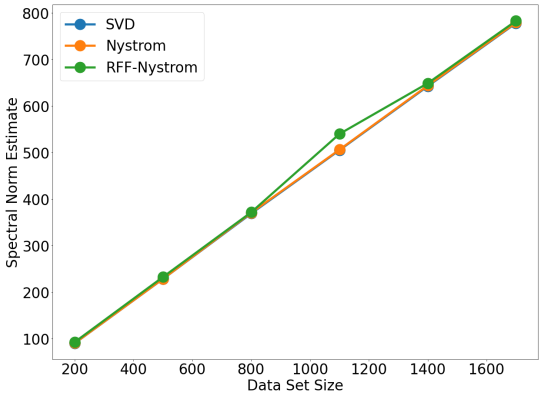
(b) Estimation Time on Crime



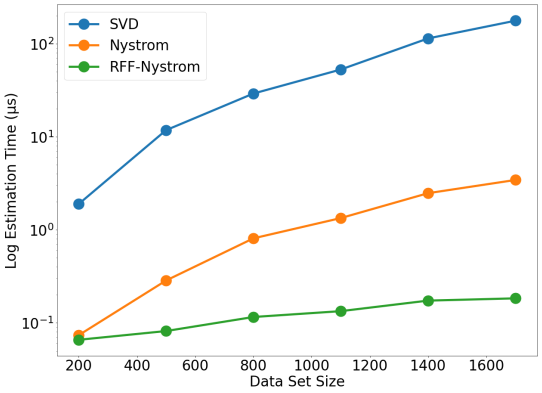
(c) Power-based Norm Estimates on Wine



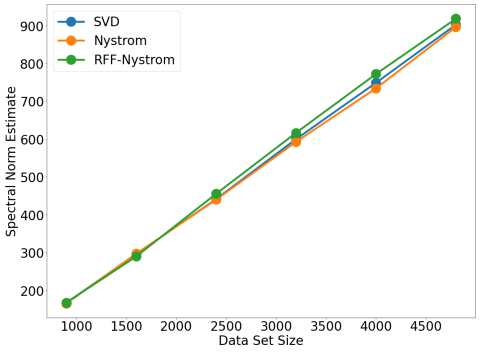
(d) Estimation Time on Wine



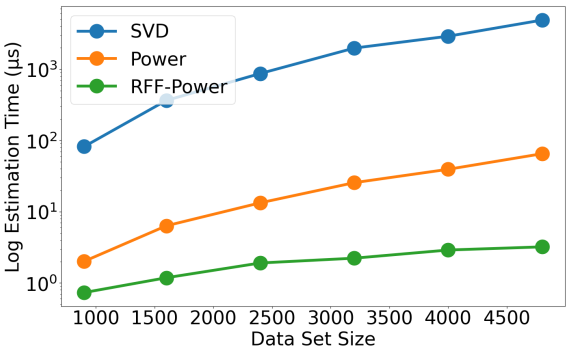
(e) Nystrom-based Norm Estimates on Crime



(f) Estimation Time on Crime

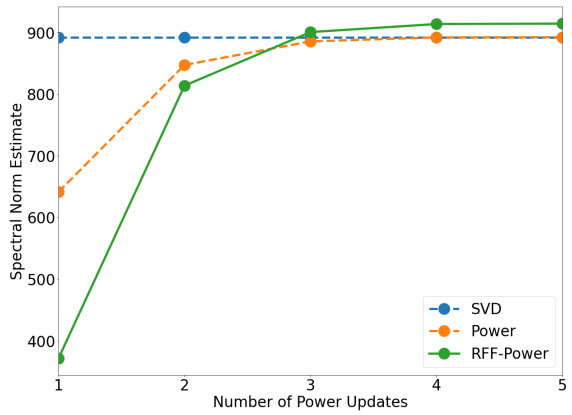


(g) Nystrom-based Norm Estimates on Wine

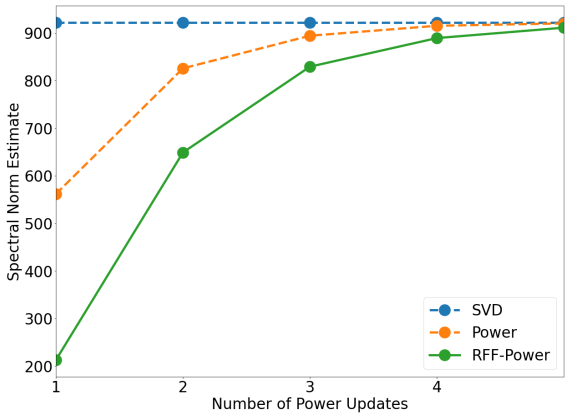


(h) Estimation Time on Wine

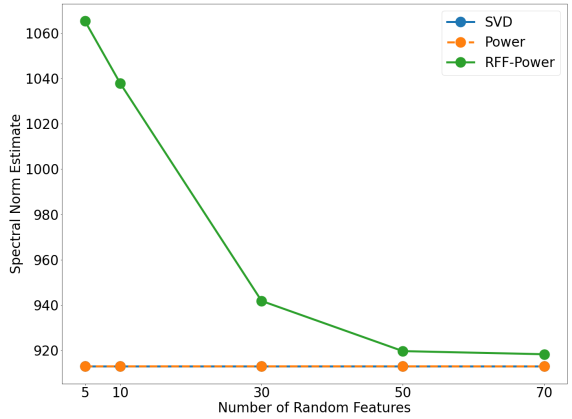
Figure 1



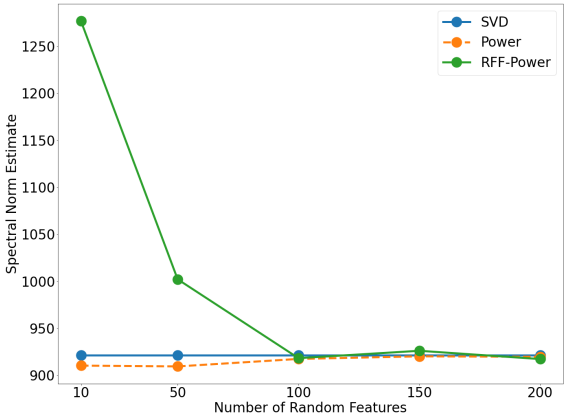
(a) Estimate vs Power Update on Crime



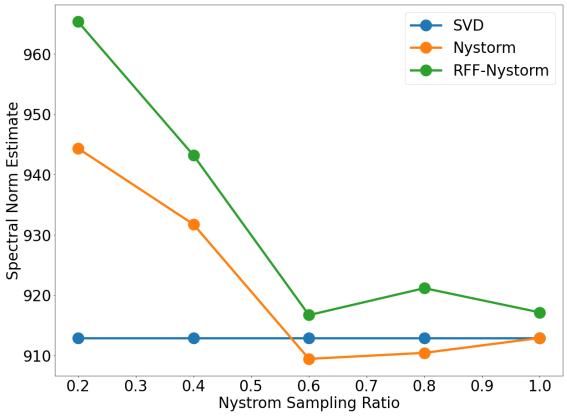
(b) Estimate vs Power Update on Wine



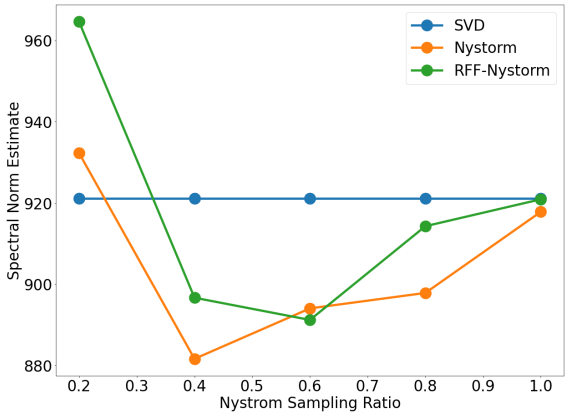
(c) Estimate vs Random Feature Number on Crime



(d) Estimate vs Random Feature Number on Wine



(e) Estimate vs Sampling Ratio on Crime



(f) Estimate vs Sampling Ratio on Wine

Figure 2: Sensitivity Analysis

4.2 Sensitivity Analysis

We also perform sensitivity analysis on different estimation methods.

Figure 2 (a-b) show the performance of Power and RFF-Power versus the number of power updates. We see both methods general converge after 5 updates, and Power converges slightly faster and more accurately than RFF-Power.

Figure 2 (c-d) show the performance of RFF-Power versus the number of random features q. We see its estimate converges to the baseline at a rate close to our theoretical prediction $\tilde{Q}(1/\sqrt{q})$. On Crime, it becomes close to baseline at $q = 50$, which is way smaller than the data size $n = 1993$; on Wine, it becomes close $q = 150$, which is also way smaller than the data size $n = 4898$. These allow RFF-Power to work with random feature matrices that are much smaller than the corresponding Gram matrices and hence gains efficiency.

Figure 2 (e-f) show the performance of Nystrom-based estimation methods versus data sampling ratio. We see both methods converge to the baseline as the ratio increases. Both convergence rates are similar, implying the proposed norm approximation does not introduce additional estimation loss for Nystrom. However, we do not observe larger variance on the performance than in other experiments. This may be because the downsampled data set no longer has standardized features to guarantee numerical stability in performance.

5 Conclusion

This paper proposes to approximate the spectral norm of a Gram matrix for n given data points using the random Fourier features of these data. Original theoretical analysis suggests an $\tilde{O}(n/\sqrt{q})$ approximation error that matches the known error of approximating kernel functions using random features, and an $\tilde{O}(\ln n/\sqrt{q})$ relative approximation error that is very close to the known spectral norm estimation error using power iteration. We then demonstrate applications of the proposed norm approximation to speed up two popular spectral norm estimation methods: power iteration and Nystrom. Experimental results on two real-world data sets demonstrate the efficacy of the accelerated methods.

Broader Impact Statement

This work proposes a novel spectral norm approximation method for Gram matrix that can be applied to significantly accelerate its estimation for scalable data analytics. There are many potential uses and therefore societal consequences of such methods, none of which we see the need to specifically highlight here.

References

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory*, pp. 458–469. Springer, 2005.

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

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Florent Benaych-Georges and Romain Couillet. Spectral analysis of the gram matrix of mixture models. *ESAIM: Probability and Statistics*, 20:217–237, 2016.

Alberto Bietti, Grégoire Mialon, Dexiong Chen, and Julien Mairal. A kernel perspective for regularizing deep neural networks. In *International Conference on Machine Learning*, pp. 664–674. PMLR, 2019.

Wuyang Chen, Xinyu Gong, and Zhangyang Wang. Neural architecture search on imagenet in four gpu hours: A theoretically inspired perspective. *arXiv preprint arXiv:2102.11535*, 2021.

Tri Dao, Christopher M De Sa, and Christopher Ré. Gaussian quadrature for kernel features. *Advances in neural information processing systems*, 30, 2017.

Tu Do and Ngoc Hoang Luong. Training-free multi-objective evolutionary neural architecture search via neural tangent kernel and number of linear regions. In *International Conference on Neural Information Processing*, pp. 335–347. Springer, 2021.

Petros Drineas, Michael W Mahoney, and Nello Cristianini. On the nyström method for approximating a gram matrix for improved kernel-based learning. *journal of machine learning research*, 6(12), 2005.

James F Epperson. *An introduction to numerical methods and analysis*. John Wiley & Sons, 2021.

Alex Gittens and Michael Mahoney. Revisiting the nystrom method for improved large-scale machine learning. In *International Conference on Machine Learning*, pp. 567–575. PMLR, 2013.

Walid Hachem, Philippe Loubaton, and J Najim. The empirical eigenvalue distribution of a gram matrix: From independence to stationarity. *arXiv preprint math/0502535*, 2005.

John T Holodnak and Ilse CF Ipsen. Randomized approximation of the gram matrix: Exact computation and probabilistic bounds. *SIAM Journal on Matrix Analysis and Applications*, 36(1):110–137, 2015.

David C Hoyle and Magnus Rattray. A statistical mechanics analysis of gram matrix eigenvalue spectra. In *International Conference on Computational Learning Theory*, pp. 579–593. Springer, 2004.

Louis Komzsik. *The Lanczos method: evolution and application*. SIAM, 2003.

Jacek Kuczyński and Henryk Woźniakowski. Estimating the largest eigenvalue by the power and lanczos algorithms with a random start. *SIAM journal on matrix analysis and applications*, 13(4):1094–1122, 1992.

Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 299–308. IEEE, 2010.

Chao Ma, Lei Wu, and E Weinan. The slow deterioration of the generalization error of the random feature model. In *Mathematical and Scientific Machine Learning*, pp. 373–389. PMLR, 2020.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322, 2010.

Marina Munkhoeva, Yermek Kapushev, Evgeny Burnaev, and Ivan Oseledets. Quadrature-based features for kernel approximation. *Advances in neural information processing systems*, 31, 2018.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008.

Kevin Roth, Yannic Kilcher, and Thomas Hofmann. Adversarial training is a form of data-dependent operator norm regularization. *Advances in Neural Information Processing Systems*, 33:14973–14985, 2020.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. *Advances in neural information processing systems*, 30, 2017.

Bernhard Schölkopf, John Shawe-Taylor, Alexander J Smola, and Robert C Williamson. Generalization bounds via eigenvalues of the gram matrix. 1999.

John Shawe-Taylor, Chris Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In *International Conference on Algorithmic Learning Theory*, pp. 23–40. Springer, 2002.

John Shawe-Taylor, Christopher KI Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51 (7):2510–2522, 2005.

Rakesh Shivanna, Bibaswan K Chatterjee, Raman Sankaran, Chiranjib Bhattacharyya, and Francis Bach. Spectral norm regularization of orthonormal representations for graph transduction. *Advances in Neural Information Processing Systems*, 28, 2015.

Taiji Suzuki. Fast generalization error bound of deep learning from a kernel perspective. In *International Conference on Artificial Intelligence and Statistics*, pp. 1397–1406. PMLR, 2018.

Martin Takác, Avleen Bijral, Peter Richtárik, and Nati Srebro. Mini-batch primal and dual methods for svms. In *International Conference on Machine Learning*, pp. 1022–1030. PMLR, 2013.

Lechao Xiao, Jeffrey Pennington, and Samuel Schoenholz. Disentangling trainability and generalization in deep neural networks. In *International Conference on Machine Learning*, pp. 10462–10472. PMLR, 2020.

Hayata Yamasaki, Sathyawageeswar Subramanian, Sho Sonoda, and Masato Koashi. Learning with optimized random features: Exponential speedup by quantum machine learning without sparsity and low-rank assumptions. *Advances in Neural Information Processing Systems*, 33:13674–13687, 2020.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. *Advances in neural information processing systems*, 25, 2012.

Felix Xinnan X Yu, Ananda Theertha Suresh, Krzysztof M Choromanski, Daniel N Holtmann-Rice, and Sanjiv Kumar. Orthogonal random features. *Advances in neural information processing systems*, 29, 2016.

Jiaru Zhang, Yang Hua, Zhengui Xue, Tao Song, Chengyu Zheng, Ruhui Ma, and Haibing Guan. Robust bayesian neural networks by spectral expectation bound regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3815–3824, 2021.