Adaptive Sampling for Low-Rank Matrix Approximation in the Matrix-Vector Product Model

Arvind K. Rathnashyam
Department of Mathematics
Rensselaer Polytechnic Institute
Troy, NY 12180, USA
rathna@rpi.edu

Nicolas Boullé
Department of Mathematics
and Theoretical Physics
University of Cambridge
Cambridge, CB3 0WA, UK
nb690@cam.ac.uk

Alex Townsend
Department of Mathematics
University of Cornell
Ithaca, NY 14853, USA
townsend@cornell.edu

Abstract

We consider the problem of low-rank matrix approximation in case the when the matrix A is accessible only via matrix-vector products and we are given a budget of $k + p$ matrix-vector products. This situation arises in practice when the cost of data acquisition is high, despite the Numerical Linear Algebra (NLA) costs being low. We create an adaptive sampling algorithm to optimally choose vectors to sample. The Randomized Singular Value Decomposition (rSVD) is an effective algorithm for obtaining the low rank representation of a matrix developed by [16]. Recently, [4] generalized the rSVD to Hilbert-Schmidt Operators where functions are sampled from non-standard Covariance Matrices when there is already prior information on the right singular vectors within the column space of the target matrix, A. In this work, we develop an adaptive sampling framework for the Matrix-Vector Product Model which does not need prior information on the matrix A. We provide a novel theoretic analysis of our algorithm with subspace perturbation theory. We extend the analysis of [22] for right singular vector approximations from the randomized SVD in the context of non-symmetric rectangular matrices. We also test our algorithm on various synthetic, real-world application, and image matrices. Furthermore, we show our theory bounds on matrices are stronger than state-of-the-art methods with the same number of matrix-vector product queries.

1 Introduction

In many real-world applications, it is often not possible to run experiments in parallel. Consider the following setting, there are a set of n inputs and m outputs, and there exists a PDE such it maps any set of inputs in $\mathbb{C}^m \to \mathbb{C}^n$. However, to run experiments, it takes hours for set up, execution, or it is expensive, e.g. aerodynamics [14], fluid dynamics [18]. Thus, after each experimental run, we want to sample a function such that in expectation, we will be exploring an area of the PDE which we have the least knowledge of. For Low-Rank Approximation the Randomized SVD, [16], has been theoretically analyzed and used in various applications. Even more recently, [2] discovered if we have prior information on the right singular vectors of A, we can modify the Covariance Matrix such that the sampled vectors are within the column space of A. They extended the theory for Randomized SVD where the covariance matrix is now a general PSD matrix. The basis of our analysis is the idea of sampling vectors in the Null-Space of the Low-Rank Approximation. This idea has been introduced recently in Machine Learning in [23] for training neural networks for sequential tasks. In a Bayesian sense, we want to maximize the expected information gain of the PDE in each iteration by sampling in the space where we have no information. This leads to the formulation of our iterative algorithm for sampling vectors for the Low-Rank Approximation. The current state of the art algorithms for low-rank matrix approximation in the matrix-vector product

Preprint. Under review.

model used a fixed covariance matrix structure. In this paper, we consider the adaptive setting where the algorithm $\mathcal{A}$ chooses a vector $\mathbf{x}^{(k)}$ with access to the previous query vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k-1)}$, the matrix-vector products $A\mathbf{x}^{(1)}, \ldots, A\mathbf{x}^{(k-1)}$, and the intermediate low-rank matrix approximations, $Q^{(k)}(Q^{(k)})^* A$, where $Q^{(k)} \triangleq \text{orth}(AV^{(k)})$ where $V^{(k)}$ is the concatenation of vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(k)}$ and $Q^{(k)} \triangleq \text{orth}(AV^{(k)})$.

Adaptive Sampling techniques for Low-Rank Matrix Approximation first appeared in CUR Matrix Decomposition in [15]. Optimial column-sampling for the CUR Matrix Decomposition received much attention as can be seen in the works [17, 10, 11]. More recently, [20] gave an algorithm for sampling the rows for CUR-Matrix Factorization and proved error bounds by induction. Similar to adaptively choosing a function, in recommender systems, the company can ask users for surveys and obtain data with high probability is a better representation of the column space of A than a random sample. Choosing the right people to give an incentivized survey (e.g. gift card upon completition) can save a company significant expenses.

Adaptively sampling vectors for matrix problems has been studied in detail in [21]. The theoretical properties of adaptively sampled matrix vector queries for estimating the minimum eigenvalue of a Wishart matrix have been studied in [7]. Their bounds are used in [1] to develop adaptive bounds for their low-rank matrix approximation method using Krylov Subspaces. To our knowledge, we are the first paper to give an algorithm for low-rank approximation in the non-symmetric matrix low-rank approximation in the matrix-vector product model. Our algorithm utilizes the SVD computaton of the low-rank approximation at each step to sample the next vector. Although there are are runtime limitations, both in theory under certain conditions and most real-world matrices, our algorithm gets the most value out of each sampled vector.

We will now clearly state our contributions.

Main Contributions.

1. We develop a novel adaptive sampling algorithm for Low-Rank Matrix Approximation problem in the matrix-vector product model which does not utilize prior information of A.
2. We provide a novel theoretical analysis which utilizes subspace perturbation theory.
3. We perform extensive experiments on matrices with various spectrums and show the effectiveness of Bayes Near-Optimal Sampling comparing to State-of-the-Art Low-Rank Matrix Approximation Algorithms in the Matrix-Vector Product Query Model.

2 Adaptive Sampling

NB: We need a good literature review and background material about rSVD here, see [Martinsson, Tropp, Acta Numerica].

In this section, we will go over the covariance matrices proposed papers and we consider choosing the optimal covariance matrix adaptively for sampling vectors. In the seminal paper by [16], the covariance matrix is given as identity matrix, $C \triangleq I$. In the generalization of the Randomized SVD, when given some prior information of the matrix, the covariance matrix is given as $C \triangleq K$ where K has some information on the right singular vectors of A (e.g. discretization of Green's Function of a PDE). Let $\tilde{V}$ be the right singular vectors of the SVD of the low-rank approximation at iteration $k-1$, then the update for the covariance matrix is given as $C^{(k+1)} \triangleq \tilde{V}_{(:,k)}\tilde{V}^*_{(:,k)}$. Throughout this paper we will only consider $C^{(0)} = I$, however using theory from [4], this can be extended to $C^{(0)} = K$ if one has some knowledge of the right singular vectors, e.g. if the matrix represents a Partial Differential Equation (PDE), having K as a discretized Green's Function for the type of PDE, e.g. elliptic, parabolic, or hyperbolic [13]. A similar algorithm can be found in [23]. It is intuitive that we want to continuously sample in the null space of the the matrix approximation we have already obtained. This ensures we are learning new information in each iteration as we don't want to sample vectors which will not learn significant unknown information about the matrix. We now give the rigourous problem definition.

Definition 1. *Given access to vector queries* $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k-1)}$ *where w.l.o.g* $\|\mathbf{v}^{(\zeta_1)}\| = 1$ *for all* $i \in [k-1]$ *and* $(\mathbf{v}^{(\zeta_1)})^* \mathbf{v}^{(j)} = \delta_{ij}$ *for all* $i, j \in [k-1] \times [k -$

1], *matrix-vector queries,* $A\mathbf{v}^{(1)}, \ldots, A\mathbf{v}^{(k-1)}$*, and intermediate low-rank approximations,* $Q^{(1)}(Q^{(1)})^* A, \ldots, Q^{(k-1)}(Q^{(k-1)})^* A$*, which requires* $k = O(\Xi)$ *vector queries to obtain a rank-*k *matrix with orthogonal columns,* Q*, such that with probability at least* $1 - \delta$ *for a* $\delta \in (0, 1)$ *such that*

$$\|A - QQ^*A\|_{\mathrm{F}} \leq (1 + \varepsilon) \min_{\substack{U \in \mathbb{C}^{m \times k} \\ U^*U = I_k}} \|A - UU^*A\|_{\mathrm{F}}$$

2.1 Algorithm

NB: We are there 2 algorithms here? We need to explain them in details underneath with motivation and intuition, also discuss benefits compared to alternatives such as fewer number of queries. State that our main objective is to use as few matvecs as possible.

The Pseudo Code for the optimal function sampling is given in Algorithm 1. For efficient updates, we frame all operations as rank-1 updates.

Algorithm 1 Near-Optimal Adaptive Sampling for Low-Rank Matrix Approximation

input: Target Matrix $A \in \mathbb{C}^{m \times n}$, target rank k, and oversampling parameter p.
output: Low-Rank Approximation $\hat{A}$ of A.

$\widehat{A}^{(0)} \leftarrow \mathbf{0}$
Form $\Omega \in \mathbb{C}^{n \times p}$ with columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, I)$
$Y^{(0)} \leftarrow A\Omega \in \mathbb{C}^{m \times p}$ and obtain the factorization $Y^{(0)} = QR$
for $t \in [k + p]$ **do**
    $\widehat{A}^{(t)} \leftarrow \widehat{A}^{(t)} + Q_{[:,t-1]} Q^*_{[:,t-1]} A$
    $C^{(t)} \leftarrow (\widehat{A}^{(t)})^+ \widehat{A}^{(t)}$
    Sample $\boldsymbol{\omega}^{(t)} \sim \mathcal{N}(\mathbf{0}, C^{(t)})$ and compute $A\boldsymbol{\omega}^{(t)}$
    $Y^{(t)} \leftarrow [Y^{(t-1)}, A\boldsymbol{\omega}^{(t)}]$ and obtain the factorization $Y^{(t)} = QR$
return: $\widehat{A}^{(k+p)}$

Algorithm 2 Power-Method Adaptive Sampling for Low-Rank Matrix Approximation

input: Target Matrix $A \in \mathbb{C}^{m \times n}$, target rank k, and oversampling parameter p.
output: Low-Rank Approximation $\hat{A}$ of A.

$\widehat{A}^{(0)} \leftarrow 0$
Sample $\boldsymbol{\omega} \in \mathbb{C}^n$ from $\mathcal{N}(\mathbf{0}, I)$
$Y^{(0)} \leftarrow A\boldsymbol{\omega} \in \mathbb{C}^{m \times p}$ and obtain the factorization $Y^{(0)} = QR$
for $t \in [k + p]$ **do**
    $\widehat{A}^{(t)} \leftarrow \widehat{A}^{(t)} + Q_{[:,t-1]} Q^*_{[:,t-1]} A$
    $\boldsymbol{\omega}^{(t)} \leftarrow$ POWER ITERATION$\big((I - QQ^*)A, \boldsymbol{\omega}, p\big)$
    $Y^{(t)} \leftarrow [Y^{(t-1)}, A\boldsymbol{\omega}^{(t)}]$ and obtain the factorization $Y^{(t)} = QR$
return: $\widehat{A}^{(k+p)}$

In Algrithm 1, we first sample a standard normal gaussian matrix which can be considered as the oversampling vectors. These oversampling vectors are used to approximate the first singular vector. This is the first vector which is *adaptively* sampled. Next, we form the low-rank approximation QQ^*A where $Q = \mathrm{orth}(A\Omega)$ where Ω is a matrix of all our adaptively chosen vector queries. From here, we choose the kth right singular vector of the SVD of the the appoximation QQ^*A. This process is then repeated for k iterations. The final low-rank approximation, QQ^*A, is then returned.

3 Theory

In this section we will first give the mathematical setup for the theoretical analysis. The proofs of all results presented in this section are deferred to the appendix.

NB: Maybe say something like "the proofs of all results in this section are deferred to the appendix". Then remove all proof statements but keep a "proof idea" paragraph detailing the main idea leading to the theorem and its main consequences. E.g.: The following result states that [...] and can be deduced from [...].

3.1 Near Optimal Function Sampling

In this section, we will discuss the theoretical motivation for adaptive sampling. Adaptive sampling in works such as [11] and [20] focus on the residual matrix, $R_t = A - Q_t Q_t^* A$. Paul et al. [20] consider the residual matrix as the matrix remaining after removing known information about A. Our first result is to formalize this statement and prove that indeed, adaptive low rank matrix approximation is at each iteration equivalent to low-rank approximation on the Residual Matrix.

Lemma 2. *Let $\Omega_+ = [\Omega, \Omega_-]$, $Y = A\Omega$, $Q = \mathrm{orth}(Y)$, and $Q_+ = \mathrm{orth}([Y, A\Omega_-])$. Finally, for shorthand, define $P_\perp = I - QQ^*$, then for $\xi \in \{2, \mathrm{F}\}$,*

$$\|A - Q_+ Q_+^* A\|_\xi = \|P_\perp A - Q_- Q_-^* P_\perp A\|_\xi \tag{1}$$

From Lemma 2, to minimize the RHS of Equation (1) we observe that Q_- is equal to the dominant left singular space of $P_\perp A$. From this result, we then see that at each iteration, the optimal vector query is the top right singular vector of $P_\perp A$. Moving forward, we then discuss how we can use the intermediate low rank approximations $Q_t Q_t^* A$ for $t \in [k+p]$ to obtain sampling vectors closer to the top right singular vector of $P_\perp A$.

3.2 Analysis of Algorithm 1

NB: This section should be extended so that a reader can understand the main results (add text between results), remove the less essential lemmas and remove proof statements.

First we will introduce a lemma for the resultant vector of sampling from $C^{(k)}$. Since our general proof technique will be an induction. We first want to understand how well we are able to approximate the first right singular vector. To do this, we must know the singular vector perturbation from the error of the low-rank matrix approximation.

Lemma 3. *Let $A \in \mathbb{C}^{m\times n}$ and Q be an orthogonal matrix representing the basis of the subspace of $Y \in \mathbb{C}^{m\times k}$. Let $C \in \mathbb{C}^{n\times n}$ such that $C \succeq \mathbf{0}$, then suppose $\Omega \in \mathbb{C}^{n\times p}$ has columns sampled i.i.d from $\mathcal{N}(\mathbf{0}, C)$. Let $Q_+ = \mathrm{orth}([Y, A\Omega])$, it then follows,*

$$\mathbf{E}\|A - Q_+ Q_+^* A\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|A - QQ^* A\|_{\mathrm{F}}^2$$

*when $p = O\Big(\frac{\|\tan\Theta(\widetilde{V}_1, C^{1/2})\|}{\varepsilon}\Big)$ where $\widetilde{V}$ is defined as in Equation (**??**).*

We will now show where we can have improvement over the randomized SVD with normal samples.

Lemma 4. *Let $A \in \mathbb{C}^{m\times n}$ and Q be an orthogonal matrix representing the basis of the subspace of $Y \in \mathbb{C}^{m\times k}$. Let $\Omega \in \mathbb{C}^{n\times p}$ such that the columns are sampled i.i.d from $\mathcal{N}(\mathbf{0}, I)$. Denote $Y_+ \triangleq \mathrm{orth}([Y, A\Omega])$ and $Q_+ \triangleq \mathrm{orth}(Y_+)$, then*

$$\mathbf{E}\|A - Q_+ Q_+^* A\|_{\mathrm{F}}^2 \leq (1+\varepsilon)\|A - QQ^* A\|_{\mathrm{F}}^2$$

when $p = O(1/\varepsilon)$.

Proof. The proof follows from a direct application of Lemma 3 and setting $C = I$. ■

Theorem 5. *Let $A \in \mathbb{C}^{m\times n}$ and let $C_i \in \mathbb{C}^{n\times n}$ for $i \in [k+p]$ be PSD covariance matrices. Then suppose $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, C_i)$ for $i \in [k+p]$. Denote $Y = \mathrm{orth}(A\Omega)$ and $Q = \mathrm{orth}(Y)$, then*

$$\|A - QQ^* A\|_{\mathrm{F}}^2 \leq \Xi$$

Lemma 6. *Let $A \in \mathbb{C}^{m\times n}$ and Q be an orthogonal matrix representing the basis of the subspace of $Y \in \mathbb{C}^{m\times k}$. Let $\Omega \in \mathbb{C}^{n\times 1}$ be the k-th right singular vector of $QQ^* A$. Denote $Y_+ \triangleq \mathrm{orth}([Y, A\Omega])$ and $Q_+ \triangleq \mathrm{orth}(Y_+)$, then*

$$\|A - Q_+ Q_+^* A\|_{\mathrm{F}}^2 \leq \|A - QQ^* A\|_{\mathrm{F}}^2 + \|\widetilde{\Sigma}_2\|_{\mathrm{F}}^2 \cdot \frac{1}{p-1} - \|A - A_k\|_2^2$$

*where $\widetilde{\Sigma}_2$ is defined as in Equation (**??**).*

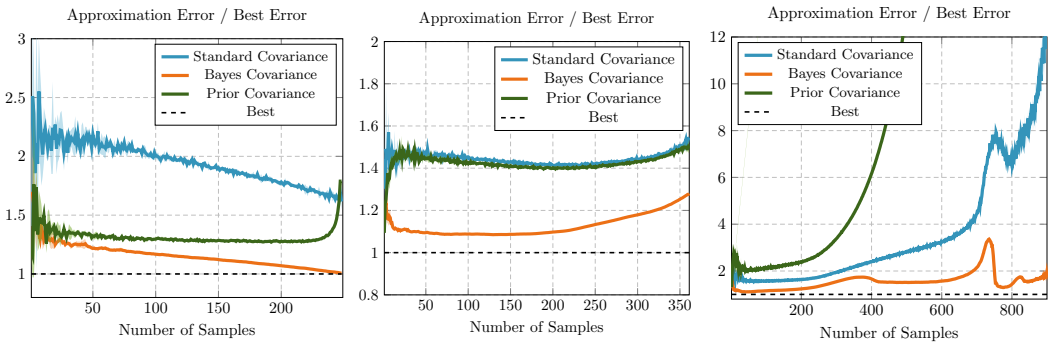

Figure 1: Low Rank Approximation for the Inverse Differential Operator given in Equation 2 Figure Left, Differential Operator Matrix `Poisson2D` [9] Figure Center, and Differential Operator Matrix `DK01R` [9] Figure Right. The experiment on the left is from [4, Figure 2]. We use the discretized Green's Function as the prior covariance matrix. In this set of experiments, we learn a low rank approximation of the inverse operator.

4 Numerical Experiments

NB: There are way too many figures here, we need to have a selection that features the main aspect of the method to discuss extensively, eventually having more experiments in Appendix. One thing we might want is comparison against Power method, computational timings. Can our method be adapted for Nystrom as well (see `https://arxiv.org/abs/2404.00960` for the general bounds)?

NB: For the tikz figures, you can generate one pdf for 3 figures by doing the following commented lines.

In this section we will test various Synthetic Matrices and Differential Operators in real-world applications with our framework and compare against the state of the art non-adaptive approaches for low-rank matrix approximation. In our first experiment we attempt to learn the discretized 250×250 matrix of the inverse of the following differential operator:

$$\mathcal{L}u = \frac{\partial^2 u}{\partial x^2} - 100 \sin(5\pi x)\, u, \qquad x \in [0, 1] \tag{2}$$

Learning the inverse operator of a PDE is equivalent to learning the Green's Function of a PDE. This has been theoretically proven for certain classes of PDEs (Linear Parabolic [3, 5]) as the inverse Differential operator is compact and there are nice theoretical properties, such as data efficiency.

In Figure 2Figure Right, note if the Covariance Matrix has eigenvectors orthogonal to the left singular vectors of A, then the randomized SVD will not perform well. Furthermore, in Figure 2Figure Left, we can note even without knowledge of the Green's Function, our method achieves lower error than with the Prior Covariance. We also test our algorithm against various Sparse Matrices in the Texas A& M Sparse Matrix Suite, [9]. In Figure 4 Figure Left, we choose a fluid dynamics problem due to its relevance in low-rank approximation [8].

Singular Value Decay. Our first synthetic matrix is developed in the following scheme:

$$A = \sum_{i=1}^{\rho} \frac{100 i^{\ell}}{n} U_{(:,i)} V^*_{(:,i)}, \quad U \in \mathbb{O}_{m,k}, V \in \mathbb{O}_{n,k} \tag{3}$$

We will experiment with linear decay, $\ell = 1$, quadratic decay, $\ell = 2$, and cubic decay, $\ell = 3$. We see adaptive sampling is as good as the Randomized SVD when there is linear decay, however, when there is quadratic decay or greater, we see that there is significant improvement when using adaptive sampling. This is corroborated in our theory, where the bounds are dependent on the ratios between the singular values. Our second synthetic matrix is developed in the following scheme:

$$A = \sum_{i=1}^{\rho} (1-\delta)^i U_{(:,i)} V^*_{(:,i)}, \quad U \in \mathbb{O}_{m,k}, V \in \mathbb{O}_{n,k} \tag{4}$$

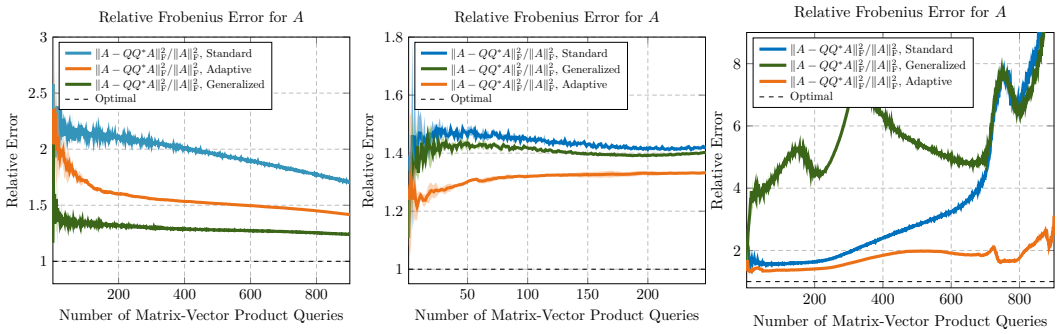

Figure 2: Results for the Power-Method based Adaptive Sampling Method described in Algorithm 2. Low Rank Approximation for the Inverse Differential Operator given in Equation 2 Figure Left, Differential Operator Matrix `Poisson2D` [9] Figure Center, and Differential Operator Matrix `DK01R` [9] Figure Right. The experiment on the left is from [4, Figure 2]. We use the discretized Green's Function as the prior covariance matrix. In this set of experiments, we learn a low rank approximation of the inverse operator.

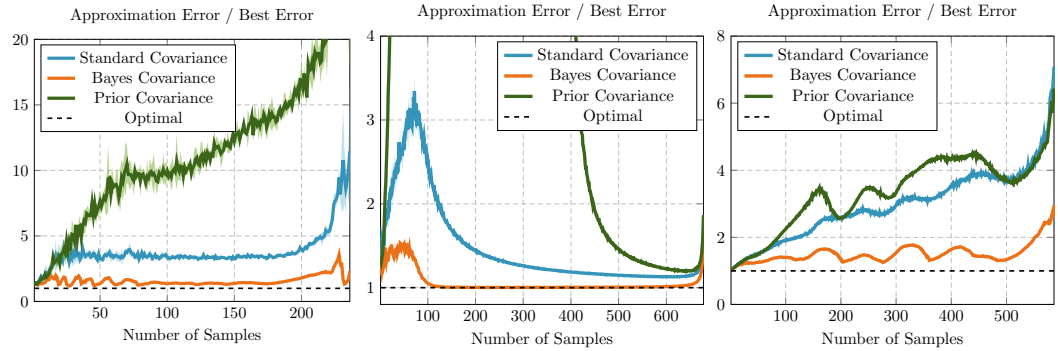

Figure 3: Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` Figure Left is from [9]. Subsequent 2D/3D Problem `fs-680-2` Figure Center is from [9]. Differential Stiffness Matrix from a Nastran Buckling Problem, `bcsstk34` Figure Right is from [9].

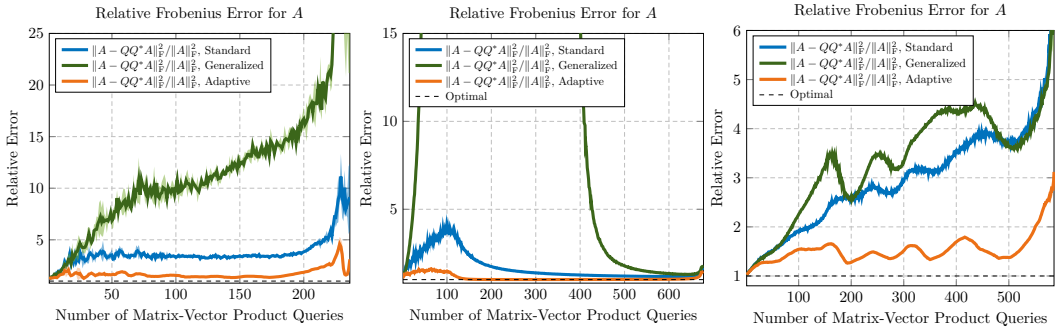

Figure 4: Results for the Power-Method based Adaptive Sampling Method described in Algorithm 2. Low Rank Approximation for a matrix for a Computational Fluid Dynamics Problem, `saylr1` Figure Left is from [9]. Subsequent 2D/3D Problem `fs-680-2` Figure Center is from [9]. Differential Stiffness Matrix from a Nastran Buckling Problem, `bcsstk34` Figure Right is from [9].

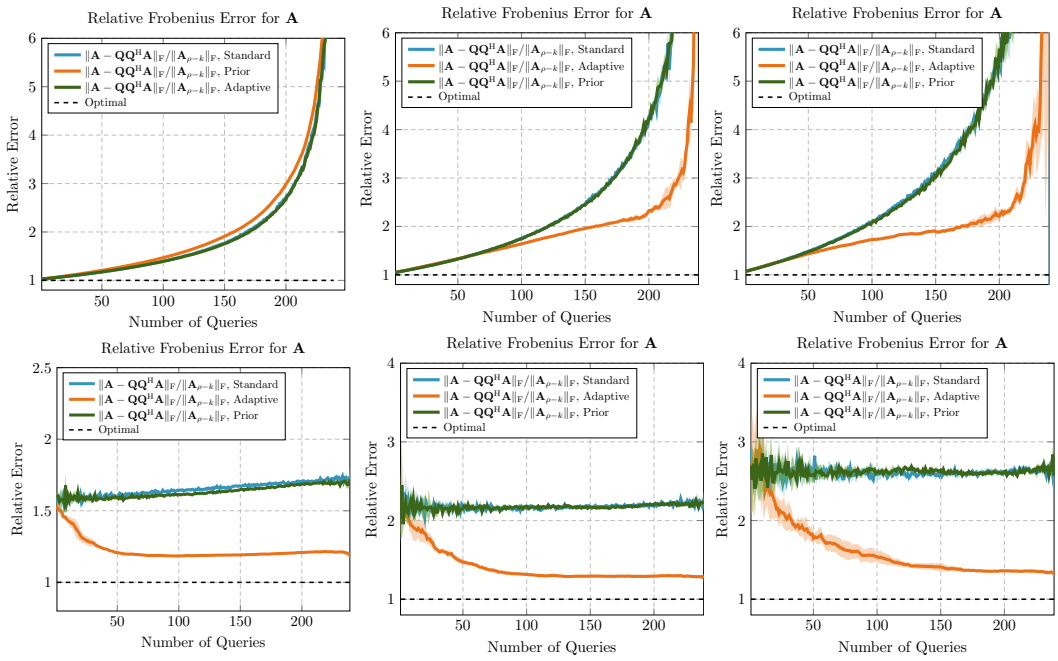

Figure 5: Low Rank Approximation for synthetic matrices with decay described in Equation 3. In *(Top Left)*, $\ell = 1$, in *(Top Center)*, $\ell = 2$, and in *(Top Right)*, $\ell = 3$. In *(Bottom Left)*, $\ell = -1$, in *(Bottom Center)*, $\ell = -2$, and in *(Bottom Right)*, $\ell = -3$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance Matrix.

We observe when there exists exponential decay of the singular values in Figure 7, adaptive sampling in accordance without scheme

5 Conclusions

We have theoretically and empirically analyzed a novel Covariance Update to iteratively construct the sampling matrix, Ω in the Randomized SVD algorithm. We introduce a new adaptive sampling framework for low-rank matrix approximation when the matrix is only accessible by matrix-vector products by giving the algorithm access to intermediate low-rank matrix approximations. Our covariance update for generating sampling vectors and functions can find use various PDE learning applications, [2, 8]. Numerical Experiments indicate without prior knowledge of the matrix, we are able to obtain superior performance to the Randomized SVD and generalized Randomized SVD with covariance matrix utilizing prior information of the PDE. Theoretically, we provide an analysis of our update extended to k-steps and show in expectation, under certain singular value decay conditions, we obtain better performance expectation.

Acknowledgments and Disclosure of Funding

The paper originated from an REU Project (give reference)... This work was supported by the Office of Naval Research (ONR), under grant N00014-23-1-2729. We thank Alex Gittens and Christopher Wang for helpful discussions.

References

[1] A. Bakshi, K. L. Clarkson, and D. P. Woodruff, *Low-rank approximation with $1/\varepsilon^3$ matrix-vector products*, in Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2022, New York, NY, USA, 2022, Association for Computing Machinery, p. 1130–1143.

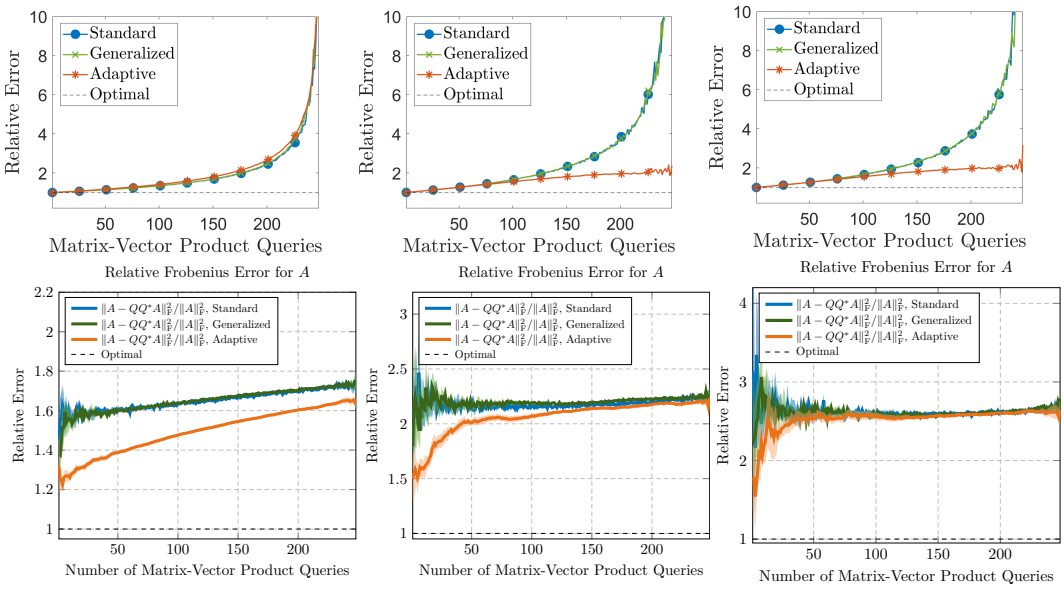

Figure 6: Results for the Power-Method based Adaptive Sampling Method desrcibed in Algorithm 2. Rank Approximation for synthetic matrices with decay described in Equation 3. In *(Top Left)*, $\ell = 1$, in *(Top Center)*, $\ell = 2$, and in *(Top Right)*, $\ell = 3$. In *(Bottom Left)*, $\ell = -1$, in *(Bottom Center)*, $\ell = -2$, and in *(Bottom Right)*, $\ell = -3$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance Matrix.

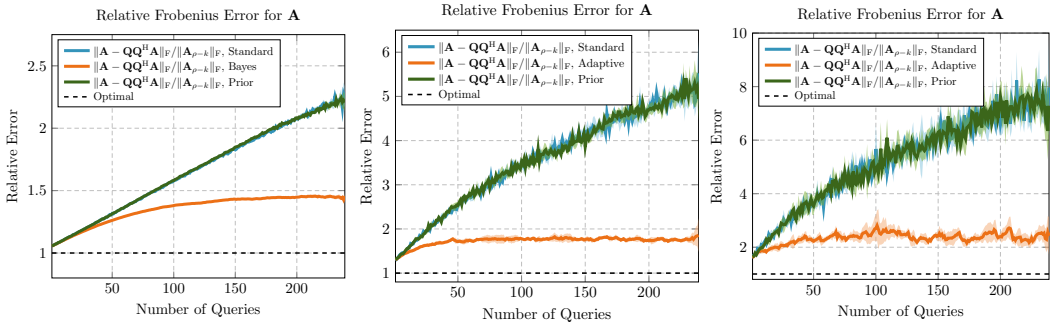

Figure 7: Low Rank Approximation for synthetic matrices with exponential decay described in Equation 4. In Figure Left, $\delta = 0.01$, in Figure Center, $\delta = 0.05$, and in Figure Right, $\delta = 0.1$. We use the Gram matrix with Gaussian Kernel and $\gamma = 0.01$ for the Prior Covariance Matrix.

[2] N. Boullé, C. J. Earls, and A. Townsend, *Data-driven discovery of green's functions with human-understandable deep learning*, Scientific Reports, 12 (2022), p. 4824.

[3] N. Boullé, S. Kim, T. Shi, and A. Townsend, *Learning green's functions associated with time-dependent partial differential equations*, The Journal of Machine Learning Research, 23 (2022), pp. 9797–9830.

[4] N. Boullé and A. Townsend, *A generalization of the randomized singular value decomposition*, in International Conference on Learning Representations, 2022.

[5] N. Boullé and A. Townsend, *Learning elliptic partial differential equations with randomized linear algebra*, Foundations of Computational Mathematics, 23 (2023), pp. 709–739.

[6] C. Boutsidis and A. Gittens, *Improved matrix algorithms via the subsampled randomized hadamard transform*, SIAM Journal on Matrix Analysis and Applications, 34 (2013), pp. 1301–1340.

[7] M. BRAVERMAN, E. HAZAN, M. SIMCHOWITZ, AND B. WOODWORTH, *The gradient complexity of linear regression*, in Conference on Learning Theory, PMLR, 2020, pp. 627–647.

[8] S. L. BRUNTON, B. R. NOACK, AND P. KOUMOUTSAKOS, *Machine learning for fluid mechanics*, Annual review of fluid mechanics, 52 (2020), pp. 477–508.

[9] T. A. DAVIS AND Y. HU, *The university of florida sparse matrix collection*, ACM Trans. Math. Softw., 38 (2011).

[10] A. DESHPANDE, L. RADEMACHER, S. S. VEMPALA, AND G. WANG, *Matrix approximation and projective clustering via volume sampling*, Theory of Computing, 2 (2006), pp. 225–247.

[11] A. DESHPANDE AND S. VEMPALA, *Adaptive sampling and fast low-rank matrix approximation*, in International Workshop on Approximation Algorithms for Combinatorial Optimization, Springer, 2006, pp. 292–303.

[12] C. ECKART AND G. YOUNG, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[13] L. C. EVANS, *Partial differential equations*, vol. 19, American Mathematical Society, 2022.

[14] H.-Y. FAN, G. S. DULIKRAVICH, AND Z.-X. HAN, *Aerodynamic data modeling using support vector machines*, Inverse Problems in Science and Engineering, 13 (2005), pp. 261–278.

[15] A. FRIEZE, R. KANNAN, AND S. VEMPALA, *Fast monte-carlo algorithms for finding low-rank approximations*, Journal of the ACM (JACM), 51 (2004), pp. 1025–1041.

[16] N. HALKO, P. G. MARTINSSON, AND J. A. TROPP, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Review, 53 (2011), pp. 217–288.

[17] S. HAR-PELED, *Low rank matrix approximation in linear time*, arXiv preprint arXiv:1410.8802, (2014).

[18] H. LOMAX, T. H. PULLIAM, D. W. ZINGG, T. H. PULLIAM, AND D. W. ZINGG, *Fundamentals of computational fluid dynamics*, vol. 246, Springer, 2001.

[19] L. MIRSKY, *Symmetric gauge functions and unitarily invariant norms*, The Quarterly Journal of Mathematics, 11 (1960), pp. 50–59.

[20] S. PAUL, M. MAGDON-ISMAIL, AND P. DRINEAS, *Column selection via adaptive sampling*, in Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds., vol. 28, Curran Associates, Inc., 2015.

[21] X. SUN, D. P. WOODRUFF, G. YANG, AND J. ZHANG, *Querying a matrix through matrix-vector products*, ACM Transactions on Algorithms (TALG), 17 (2021), pp. 1–19.

[22] R.-C. TZENG, P.-A. WANG, F. ADRIAENS, A. GIONIS, AND C.-J. LU, *Improved analysis of randomized svd for top-eigenvector approximation*, in Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, G. Camps-Valls, F. J. R. Ruiz, and I. Valera, eds., vol. 151 of Proceedings of Machine Learning Research, PMLR, 28–30 Mar 2022, pp. 2045–2072.

[23] S. WANG, X. LI, J. SUN, AND Z. XU, *Training networks in null space of feature covariance for continual learning*, in Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2021, pp. 184–193.

A Proofs

In this section we give proofs for results we deferred from the main text.

A.1 Distribution of $V^*\Omega$

We devote this section to the matrix $V^*\Omega$. We will derive various concentration inequalities which allow us to give the main theorems.

Lemma 7. *Let $\Omega = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_\ell] \in \mathbb{C}^{n\times\ell}$ such that $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, C_i)$ and $C_i \succ 0$ for all $i \in [\ell]$. Then, let $V \in \mathbb{O}_{n\times k}$, it then follows the jth column of $V^*\Omega$ is distributed as $\mathcal{N}(\mathbf{0}, K_j)$ where $K_j = \operatorname{diag}(V^*C_jV)$ for all $j \in [\ell]$.*

Proof. The matrix $V^*\Omega$ can be decomposed as follows,

$$V^*\Omega = \begin{bmatrix} \mathbf{v}_1^*\boldsymbol{\omega}_1 & \cdots & \mathbf{v}_1^*\boldsymbol{\omega}_\ell \\ \vdots & \ddots & \vdots \\ \mathbf{v}_k^*\boldsymbol{\omega}_1 & \cdots & \mathbf{v}_k^*\boldsymbol{\omega}_\ell \end{bmatrix} \tag{5}$$

Let $\mathcal{D} = \mathcal{N}(\mathbf{0}, I)$. We will first show that the entries are Gaussian, for any $i \in [k]$ and $j \in [\ell]$, we have for $\mathbf{x} \sim \mathcal{D}$,

$$\mathbf{v}_i^* \boldsymbol{\omega}_j = \mathbf{v}_i^* C_j^{1/2} \mathbf{x} = \sum_{k \in [n]} \mathbf{v}_i^* \mathbf{u}_k \sqrt{\lambda_k(C_j)} x_k \tag{6}$$

In the above, we have that each $[V^* \Omega]_{i,j}$ is Gaussian for $(i,j) \in [k] \times [\ell]$ as a linear combination of Gaussians is Gaussian. We will calculate the mean and covariance. We then have for any $(i,j) \in [k] \times [\ell]$,

$$\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j \sim \mathcal{N}(\mathbf{0}, C_j)} [\mathbf{v}_i^* \boldsymbol{\omega}_j] = \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}} [\mathbf{v}_i^* C^{1/2} \mathbf{x}] = \mathbf{v}_i^* C^{1/2} \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}} [\mathbf{x}] \tag{7}$$

$$= \sum_{p \in [n]} \mathbf{v}_i^* \mathbf{u}_j \sqrt{\lambda_p(C_j)} \mathop{\mathbf{E}}_{\mathbf{x} \sim \mathcal{D}} [x_p] = 0 \tag{8}$$

Now we calculate the covariance matrix. Let $\mathbf{v} \in \mathbb{S}^{n-1}$, then we have for any $(i, i', j, j') \in [k] \times [k] \times [\ell] \times [\ell]$, we have for the non-diagonal elements when $i \neq j$,

$$\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j \sim \mathcal{N}(\mathbf{0}, C_j)} \mathop{\mathbf{E}}_{\boldsymbol{\omega}_{j'} \sim \mathcal{N}(\mathbf{0}, C_{j'})} \left[(\mathbf{v}_i^* \boldsymbol{\omega}_j - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, C_j)} [\mathbf{v}_i^* \boldsymbol{\omega}_j])(\mathbf{v}_{i'}^* \boldsymbol{\omega}_{i'} - \mathop{\mathbf{E}}_{\boldsymbol{\omega}_{j'} \sim \mathcal{N}(\mathbf{0}, C_{j'})} [\mathbf{v}_{i'}^* \boldsymbol{\omega}_{j'}]) \right] \tag{9}$$

$$= \mathop{\mathbf{E}}_{\boldsymbol{\omega}_j \sim \mathcal{N}(\mathbf{0}, C_j)} \mathop{\mathbf{E}}_{\boldsymbol{\omega}_{j'} \sim \mathcal{N}(\mathbf{0}, C_{j'})} [(\mathbf{v}_i^* \boldsymbol{\omega}_j)(\mathbf{v}_{i'}^* \boldsymbol{\omega}_{j'})] \tag{10}$$

$$\overset{(\zeta_1)}{=} \mathop{\mathbf{E}}_{\mathbf{x}_j \sim \mathcal{D}} \left(\mathop{\mathbf{E}}_{\mathbf{x}_{j'} \sim \mathcal{D}} [(\mathbf{v}_{i'}^* C_{j'} \mathbf{x}_{j'})(\mathbf{v}_i^* C_j \mathbf{x}_j) \mid \mathbf{v}_i^* C_j \mathbf{x}_j] \right) = 0 \tag{11}$$

In the above, (ζ_1) follows from the law of total expectation. By conditioning on $\mathbf{v}_i^* \boldsymbol{\omega}_j$, we are able to obtain zero covariance even if $C_{j'}$ is dependent on C_j. For the diagonal covariance elements, we have

$$\mathop{\mathbf{E}}_{\boldsymbol{\omega}_j \sim \mathcal{N}(\mathbf{0}, C)} [(\mathbf{v}_i^* \boldsymbol{\omega}_j - \mathbf{E}[\mathbf{v}_i^* \boldsymbol{\omega}_j])^2] = \mathop{\mathbf{E}}_{\boldsymbol{\omega}_k \sim \mathcal{N}(\mathbf{0}, C)} [\mathbf{v}_i^* \boldsymbol{\omega}_j \boldsymbol{\omega}_j^* \mathbf{v}_i] = \mathbf{v}_i^* C_j \mathbf{v}_i \tag{12}$$

We thus have for all $(i, i') \in [k] \times [k]$,

$$[K_j]_{ii'} = \begin{cases} \mathbf{v}_i^* C_j \mathbf{v}_i & i = i' \\ 0 & i \neq i' \end{cases} \tag{13}$$

Our proof is complete. ■

Lemma 8. *Let $\Omega = [\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_\ell] \in \mathbb{C}^{n \times \ell}$ such that $\omega_i \sim \mathcal{N}(\mathbf{0}, C_i)$ and $C_i \succ 0$ for all $i \in [\ell]$. Suppose $V = \begin{bmatrix} V_k^* \\ V_{k,\perp}^* \end{bmatrix} \in \mathbb{O}_{n \times n}$ and fix $\Sigma = \begin{bmatrix} \Sigma_k & \\ & \Sigma_{k,\perp} \end{bmatrix}$, then with high probability*

$$\|\Sigma_{k,\perp} V_{k,\perp}^* \Omega\|_{\mathrm{F}}^2 \leq \Xi \tag{14}$$

Proof. From Lemma 7, we have that jth column of $V_{k,\perp}^* \Omega$ is sampled from $\mathcal{N}(\mathbf{0}, \operatorname{diag}(V_{k,\perp}^* C_j V_{k,\perp}))$.

$$\|\Sigma_{k,\perp} V_{k,\perp}^* \Omega\|_{\mathrm{F}}^2 \leq \Xi$$

■

Lemma 9. *Let $\Omega = [\mathbf{w}, \ldots, \boldsymbol{\omega}_\ell] \in \mathcal{C}^{n \times \ell}$ such that $\boldsymbol{\omega}_i \sim \mathcal{N}(\mathbf{0}, C_i)$ and $C_i \succ 0$ for all $i \in [\ell]$. Suppose $V = \begin{bmatrix} V_k^* \\ V_{k,\perp}^* \end{bmatrix} \in \mathbb{O}_{n \times n}$ and fix $\Sigma = \begin{bmatrix} \Sigma_k & \\ & \Sigma_{k,\perp} \end{bmatrix}$, then with high probability*

$$\|(V_k^* \Omega)^+\|_{\mathrm{F}} \leq \Xi \tag{15}$$

Proof. For shorthand let $V_k^* \Omega = \Omega_1$, then we have

$$\|\Omega_1^+\|_{\mathrm{F}}^2 = \operatorname{Tr}\big((\Omega_1^+)^* \Omega_1^+\big) \tag{16}$$

■

A.2 Proof of Lemma 3

Proof. Our first key note is that adaptive low rank matrix approximation is equivalent to performing one sample of the generalized rSVD to the residual $A - QQ^*A$. Let $Q = \mathrm{orth}(Y)$ and $Q_+ = \mathrm{orth}([Y, A\Omega])$ for a sample matrix Ω, and $Q_-Q_-^* = Q_+Q_+^* - QQ^*$, then we have

$$(A - QQ^*A) - Q_-Q_-^*(A - QQ^*A) = A - QQ^*A - Q_-Q_-^*A = A - Q_+Q_+^*A \tag{17}$$

We first give the factorization $A \triangleq (I - QQ^*)A + QQ^*A \triangleq P_\perp A + PA$, it then follows $Q_- = \mathrm{orth}((I - QQ^*)A\Omega) \overset{\text{def}}{=} \mathrm{orth}(P_\perp A\Omega) \triangleq \mathrm{orth}(\widetilde{Y}_-)$. Then, we factorize the residual term as follows,

$$P_\perp A \triangleq (I - QQ^*)A \triangleq \begin{matrix} k & n-k \\ \end{matrix} \begin{bmatrix} \widetilde{U}_k & \widetilde{U}_{k,\perp} \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix} \begin{bmatrix} \widetilde{\Sigma}_k & \\ & \widetilde{\Sigma}_{k,\perp} \end{bmatrix} \begin{bmatrix} \widetilde{V}_k^* \\ \widetilde{V}_{k,\perp}^* \end{bmatrix} \begin{matrix} k \\ n-k \end{matrix} \tag{18}$$

Defining $\widetilde{\Omega}_k \triangleq \widetilde{V}_k^*\Omega$ and $\widetilde{\Omega}_{k,\perp} \triangleq \widetilde{V}_{k,\perp}^*\Omega$. We then have,

$$\mathbf{E}\|A - Q_+Q_+^*A\|_{\mathrm{F}} \overset{(17)}{=} \mathbf{E}\|P_\perp A - Q_-Q_-^*P_\perp A\|_{\mathrm{F}} \tag{19}$$

$$\overset{(\zeta_1)}{\leq} \left(\mathbf{E}\|P_\perp A - Q_-Q_-^*P_\perp A\|_{\mathrm{F}}^2\right)^{1/2} \tag{20}$$

$$\overset{(\zeta_2)}{\leq} \left(\|\widetilde{\Sigma}_{k,\perp}\|_{\mathrm{F}}^2 + \mathbf{E}\|\widetilde{\Sigma}_{k,\perp}\widetilde{\Omega}_{k,\perp}\widetilde{\Omega}_k^+\|_{\mathrm{F}}^2\right)^{1/2} \tag{21}$$

where (ζ_1) follows from Jensen's Inequality and (ζ_2) follows from [16, Theorem 9.1], which holds from noting that $Q_- = \mathrm{orth}(P_\perp A\Omega_-)$. Term by term, we have

$$\|\widetilde{\Sigma}_{k,\perp}\|_{\mathrm{F}}^2 = \|P_\perp A\|_{\mathrm{F}}^2 - \|P_\perp A\|_2^2 = \|A - QQ^*A\|_{\mathrm{F}}^2 - \|A - QQ^*A\|_2^2 \overset{(\zeta_3)}{\leq} \|A - QQ^*A\|_{\mathrm{F}}^2 - \|A - A_k\|_2^2 \tag{22}$$

where in (ζ_3) we used the following,

$$\|A - QQ^*A\|_2^2 \geq \min_{\substack{P \in \mathbb{C}^{m\times m} \\ \mathrm{rank}(P)=k \\ P^2=P}} \|A - PA\|_2^2 \geq \min_{\substack{B \in \mathbb{C}^{m\times n} \\ \mathrm{rank}(B)=k}} \|A - B\|_2^2 \overset{(\zeta_4)}{=} \|A - A_k\|_2^2 \tag{23}$$

where (ζ_4) follows from the Eckart-Young-Mirsky Theorem [12, 19]. An immediate consequence of Equation (23), is that,

$$\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|A - QQ^*A\|_{\mathrm{F}}^2 + \|\widetilde{\Sigma}_{k,\perp}\widetilde{\Omega}_{k,\perp}\widetilde{\Omega}_k^+\|_{\mathrm{F}}^2 - \|A - A_k\|_2^2 \tag{24}$$

We thus have, for any algorithm we can always improve the approximation by $\|A - A_k\|_2^2$ with some approximation error under the condition that $\mathrm{rank}(Q) = k$ and $\mathrm{rank}(Q_+) = k + p$. Consider the case when $Q = \mathbf{0}$, then we recover the HMT bounds. Expanding out the error term, we have

$$\mathbf{E}\|\widetilde{\Sigma}_{k,\perp}\widetilde{\Omega}_{k,\perp}\widetilde{\Omega}_k^+\|_{\mathrm{F}}^2 = \mathbf{E}\|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}X(\widetilde{V}_k^*C^{1/2}X)^+\|_{\mathrm{F}}^2 \tag{25}$$

$$= \mathbf{E}\left(\mathbf{E}\left[\|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}X(\widetilde{V}_k^*C^{1/2}X)^+\|_{\mathrm{F}}^2 \mid \widetilde{V}_k^*C^{1/2}X\right]\right) \tag{26}$$

$$\overset{(\zeta_5)}{=} \|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}\|_{\mathrm{F}}^2 \cdot \mathbf{E}\|(\widetilde{V}_k^*C^{1/2}X)^+\|_{\mathrm{F}}^2 \tag{27}$$

where (ζ_5) follows from Lemma 11 as X is a standard Gaussian matrix. We then have from [5],

$$\mathbf{E}\|\Omega_k^+\|_{\mathrm{F}}^2 = \mathbf{E}\,\mathrm{Tr}((\Omega_k^+)^*\Omega_k^+) = \mathbf{E}\,\mathrm{Tr}((\Omega_k\Omega_k^*)^{-1}) = \mathrm{Tr}(\mathbf{E}[(\Omega_k\Omega_k^*)^{-1}]) = \frac{\mathrm{Tr}(K^{-1})}{p-2} \tag{28}$$

where from Lemma 7, we have $K = \mathrm{diag}(\widetilde{V}_k^*C\widetilde{V}_k)$. From which we obtain the claimed bound,

$$\mathbf{E}\|A - Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|[A - QQ^*A]_{k,\perp}\|_{\mathrm{F}}^2 + \frac{1}{p-2}\|\widetilde{\Sigma}_{k,\perp}\widetilde{V}_{k,\perp}^*C^{1/2}\|_{\mathrm{F}}^2 \cdot \mathrm{Tr}([\mathrm{diag}(\widetilde{V}_k^*C\widetilde{V}_k)]^{-1})$$

$$\leq \|[A - QQ^*A]_{k,\perp}\|_{\mathrm{F}}^2 + \frac{1}{p-2} \cdot \|\widetilde{\Sigma}_{k,\perp}\|_{\mathrm{F}}^2\|\tan\Theta(\widetilde{V}_k, C^{1/2})\|_2^2 \tag{29}$$

$$\overset{(\zeta_1)}{\leq} (1+\varepsilon)\|[A-QQ^*A]_{k,\perp}\|_{\mathrm{F}}^2 \tag{30}$$

In the above, (ζ_1) holds when,

$$p \geq 2 + \frac{\|\tan\Theta(\widetilde{V}_k, C^{1/2})\|_2^2}{\epsilon}$$

Our proof is complete. ■

A.3 Proof of Theorem 5

Proof. From Lemma 5.5 in [6], we have for any $\xi \in \{2, \mathrm{F}\}$,

$$\|A-QQ^*A\|_\xi^2 \leq \|\Sigma_{k,\perp}V_{k,\perp}^*\Omega(V_k^*\Omega)^+\|_\xi^2 + \|\Sigma_{k,\perp}\|_\xi^2 \tag{31}$$

$$\leq \|\Sigma_{k,\perp}V_{k,\perp}^*\Omega\|_\xi^2\|(V_k^*\Omega)^+\|_2^2 + \|\Sigma_{k,\perp}\|_\xi^2 \tag{32}$$

$$\leq \|\Sigma_{k,\perp}\|_\xi^2\big(1 + \|V_{k,\perp}^*\Omega\|_2^2\|(V_k^*\Omega)^+\|_2^2\big) \tag{33}$$

■

A.4 Proof of Lemma 6

Proof. Recall for a vector $\boldsymbol{\omega}$, we have

$$\|A-Q_+Q_+^*A\|_{\mathrm{F}}^2 \leq \|A-QQ^*A\|_{\mathrm{F}}^2 + \|\widetilde{\Sigma}_{k,\perp}\|_2^2\tan^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}) - \|A-QQ^*A\|_2^2 \tag{34}$$

It becomes clear we want to adaptively choose $\boldsymbol{\omega}$, that reduces $\tan\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega})$. This inequality is sharp when $Q = U_k$, then choosing $\boldsymbol{\omega} = \mathbf{v}_{k+1}$ is optimal. The natural approach is power iteration. Now, consider $p = 2$, we have,

$$\boldsymbol{\omega}_{k,\perp} = ((I-QQ^*)A)^*(I-QQ^*)A\mathbf{v} = A^*(I-QQ^*)A\mathbf{v} = A^*A\mathbf{v} - A^*QQ^*QQ^*A\mathbf{v} \tag{35}$$

Recall at any iteration t, we have QQ^*A, and thus the second term is known. The first term on the other hand requires two matrix-vector multiplications. Let $\mathbf{v}_i$ for $i \in [n]$ represent the right singular vectors of $(I-QQ^*)A$. Then, we have for some coefficients, α_i for $i \in [n]$ such that $\boldsymbol{\omega} = \sum_{i\in[n]}\alpha_i\widetilde{\mathbf{v}}_i$. We then, have $\boldsymbol{\omega}_{k,\perp} = A^*(I-QQ^*)A\boldsymbol{\omega} = \sum_{i\in[n]}\alpha_i\lambda_i(A^*(I-QQ^*)A)\widetilde{\mathbf{v}}_i$. Then we have,

$$\cos^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}_p) = \alpha_k^2\lambda_1^p(A^*(I-QQ^*)A) \tag{36}$$

and for $\sin$ we have

$$\sin^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}_p) = \sum_{i=2}^n\alpha_i^2\lambda_i(A^*(I-QQ^*)A) \leq \left(\sum_{i=2}^n\alpha_i^2\right)\lambda_2(A^*(I-QQ^*)A) \tag{37}$$

Combining our upper bound for $\sin$ and lower bound for $\cos$, we have

$$\tan^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}_p) \leq \left(\frac{\lambda_2(A^*(I-QQ^*)A)}{\lambda_1(A^*(I-QQ^*)A)}\right)^p\tan^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}) \tag{38}$$

We now bound the top eigenvalue ratio of $A^*(I-QQ^*)A$,

$$\lambda_1(A^*(I-QQ^*)A) = \lambda_1(A^*(I-U_kU_k^*+U_kU_k^*-QQ^*)A) \tag{39}$$

$$\geq \lambda_1(A^*A-A_k^*A_k) + \lambda_n(A^*(U_kU_k^*-QQ^*)A) \tag{40}$$

$$\geq \lambda_1(A^*A-A_k^*A_k) - \lambda_1(A^*A)\|U_kU_k^*-QQ^*\|_2 \tag{41}$$

and from the similar idea, we obtain the following upper bound from Weyl's Inequality,

$$\lambda_2(A^*(I-QQ^*)A) \leq \lambda_2(A^*A-A_k^*A_k) + \lambda_1(A^*A)\|U_kU_k^*-QQ^*\|_2 \tag{42}$$

Combining our previous two estimates, we obtain,

$$\tan^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}_p) \leq \left(\frac{\lambda_{k+1}(A^*A)+\lambda_1(A^*A)\|U_kU_k^*-QQ^*\|_2}{\lambda_1(A^*A)-\lambda_1(A^*A)\|U_kU_k^*-QQ^*\|_2}\right)^p\tan^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}) \tag{43}$$

$$\leq \left(\frac{\lambda_{k+1}(A^*A)}{\lambda_1(A^*A)}\right)^p\left(\frac{1+\epsilon}{1-\epsilon}\right)^p\tan^2\Theta(\widetilde{\mathbf{v}}_k,\boldsymbol{\omega}) \tag{44}$$

Our proof is complete.

■

B Probability Theory

Lemma 10. *Sample $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, I)$. Then with probability at least $0.8 - \delta/2$, for any $\mathbf{v} \in \mathcal{S}^{n-1}$, we have*

$$\tan\Theta(\mathbf{v}, \mathbf{x}) \leq 128n\log(6) + 128\log(4/\delta)$$

Proof. Let C be an undetermined positive constant, we then have the following decomposition,

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{\tan^2\Theta(\mathbf{v}, \mathbf{x}) \geq t\right\} = \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{\sin^2\Theta \geq t\cos^2\Theta\right\} \tag{45}$$

$$= \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{\cos^2\Theta \leq \frac{1}{t+1}\right\} \leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{|\mathbf{x}\cdot\mathbf{v}| \leq \|\mathbf{x}\|\sqrt{\frac{1}{t+1}}\right\} \tag{46}$$

$$= \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{|\mathbf{x}\cdot\mathbf{v}| \leq \|\mathbf{x}\|\sqrt{\frac{1}{t+1}}, \|\mathbf{x}\| \leq C\right\} + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{|\mathbf{x}\cdot\mathbf{v}| \leq \|\mathbf{x}\|\sqrt{\frac{1}{t+1}}, \|\mathbf{x}\| \geq C\right\} \tag{47}$$

$$\leq \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{|\mathbf{x}\cdot\mathbf{v}| \leq C\sqrt{\frac{1}{t+1}}\right\} + \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\{\|\mathbf{x}\| \geq C\} \tag{48}$$

We now make a ε-net argument. Let $\mathcal{N}$ be a ε-net of $\mathbb{S}^{n-1}$. Then let $\mathbf{v}^* = \arg\max_{\|\mathbf{v}\|=1}|\mathbf{x}\cdot\mathbf{v}|$ and let $\mathbf{u} \in \mathcal{N}$ such that $\|\mathbf{u} - \mathbf{v}^*\| \leq \varepsilon$, then we have

$$|\mathbf{x}\cdot\mathbf{u}| \geq |\mathbf{x}\cdot\mathbf{v}^*| - |\mathbf{x}\cdot\mathbf{v}^* - \mathbf{x}\cdot\mathbf{u}| \geq (1-\varepsilon)\|\mathbf{x}\| \tag{49}$$

From which we obtain, $\|\mathbf{x}\| \leq (1-\varepsilon)^{-1}\arg\max_{\mathbf{u}\in\mathcal{N}}|\mathbf{x}\cdot\mathbf{u}|$. Then, from a union bound, we have,

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\{\|\mathbf{x}\| \geq C\} \leq \left(\frac{3}{\varepsilon}\right)^n \mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\{|\mathbf{x}\cdot\mathbf{v}| \geq C(1-\varepsilon)\} \leq 6^n\exp\left(-\frac{C^2}{8}\right) \leq \frac{\delta}{2} \tag{50}$$

Suppose we choose $\varepsilon = 1/2$, then the above probabilistic condition is satisfied when

$$C^2 \geq 8\log(4/\delta) + 8n\log(6) \tag{51}$$

We then have,

$$\mathop{\mathbf{Pr}}_{\mathbf{x}\sim\mathcal{D}}\left\{|\mathbf{x}\cdot\mathbf{v}| \leq C\sqrt{\frac{1}{t+1}}\right\} \leq 0.2 \tag{52}$$

The above probabilistic condition holds when $t \geq 16C^2$. We then obtain with probability $0.8 - \delta/2$,

$$\tan\Theta(\mathbf{v}, \mathbf{x}) \leq 128n\log(6) + 128\log(4/\delta) \tag{53}$$

Our proof is complete. ■

Lemma 11 (Proposition 10.1 [16])**.** *Fix matrices $S \in \mathbb{R}^{K\times N}$ and $T \in \mathbb{R}^{N\times L}$, then for a matrix G such that elements of G are sampled i.i.d from $\mathcal{N}(0,1)$, then*

$$\left(\mathbf{E}\|SGT\|_{\mathrm{F}}^2\right)^{1/2} = \|S\|_{\mathrm{F}}\|T\|_{\mathrm{F}} \tag{54}$$

Proof. The proof is a simple calculation.

$$\mathbf{E}\|SGT\|_{\mathrm{F}}^2 = \sum_{i\in[K]}\sum_{j\in[L]}\sum_{(k_1,k_2)\in[N]\times[N]} S_{i,k_1}^2 T_{k_2,j}^2\,\mathbf{E}\left[G_{k_1,k_2}^2\right] \tag{55}$$

$$= \sum_{i\in[K]}\sum_{j\in[L]}\sum_{(k_1,k_2)\in[N]\times[N]} S_{i,k_1}^2 T_{k_2,j}^2 \tag{56}$$

$$= \|S\|_{\mathrm{F}}^2\|T\|_{\mathrm{F}}^2 \tag{57}$$

Taking the square root of both sides completes the proof. ■

Lemma 12. *Fix matrices $S \in \mathbb{R}^{K\times N}$ and $T \in \mathbb{R}^{N\times L}$, then for a matrix G such that elements of G are sampled i.i.d from $\mathcal{N}(0,1)$, then with probability exceeding $1-\delta$,*

$$\|SGT\|_{\mathrm{F}}^2 \leq 2\|S\|_{\mathrm{F}}\|T\|_{\mathrm{F}}\log(N^2/\delta) \tag{58}$$

Proof. The proof follows similarly as Lemma 11.

$$\|SGT\|_{\mathrm{F}}^2 = \sum_{i\in[K]}\sum_{j\in[L]}\sum_{(k_1,k_2)\in[N]\times[N]} S_{i,k_1}^2 T_{k_2,j}^2 G_{k_1,k_2}^2 \tag{59}$$

$$\leq \sum_{i\in[K]}\sum_{j\in[L]}\sum_{(k_1,k_2)\in[N]\times[N]} S_{i,k_1}^2 T_{k_2,j}^2 \max_{(k_1,k_2)\in[N]\times[N]} G_{k_1,k_2}^2 \tag{60}$$

We now bound the maximum Gaussian over a finite sample.

$$\mathbf{Pr}\left\{\max_{(k_1,k_2)\in[N]\times[N]} G_{k_1,k_2}^2 \geq t\right\} = \mathbf{Pr}\left\{\max_{(k_1,k_2)\in[N]\times[N]} |G_{k_1,k_2}| \geq \sqrt{t}\right\}$$

$$\leq \frac{\sqrt{2}N^2}{\sqrt{\pi}}\int_{\sqrt{t}}^{\infty} e^{-x^2/2}dx \leq \frac{\sqrt{2}N^2}{\sqrt{\pi}}\int_{\sqrt{t}}^{\infty}\frac{xe^{-x^2/2}}{\sqrt{t}}dx = \frac{\sqrt{2}N^2}{\sqrt{\pi}}e^{-t/2} \leq \delta \tag{61}$$

Then, from some algebra, we obtain with probability exceeding $1-\delta$,

$$\|SGT\|_{\mathrm{F}}^2 \leq 2\log(N^2/\delta)\|S\|_{\mathrm{F}}^2\|T\|_{\mathrm{F}}^2 \tag{62}$$

The proof is complete. ■

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.