Privacy-preserving Sparse Generalized Eigenvalue Problem

Lijie Hu KAUST
Zihang Xiang KAUST
Jiabin Liu Beijing Institute of Technology
Di Wang KAUST

Abstract

In this paper we study the (sparse) Generalized Eigenvalue Problem (GEP), which arises in a number of modern statistical learning models, such as principal component analysis (PCA), canonical correlation analysis (CCA), Fisher's discriminant analysis (FDA) and sliced inverse regression (SIR). We provide the first study on GEP in the differential privacy (DP) model under both deterministic and stochastic settings. In the low dimensional case, we provide a ρ-Concentrated DP (CDP) method namely DP-Rayleigh Flow and show if the initial vector is close enough to the optimal vector, its output has an ℓ_2-norm estimation error of $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ (under some mild assumptions), where d is the dimension and n is the sample size. Next, we discuss how to find such a initial parameter privately. In the high dimensional sparse case where $d \gg n$, we propose the DP-Truncated Rayleigh Flow method whose output could achieve an error of $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$ for various statistical models, where s is the sparsity of the underlying parameter. Moreover, we show that these errors in the stochastic setting are optimal up to a factor of $\text{Poly}(\log n)$ by providing the lower bounds of PCA and SIR under statistical setting and in the CDP model. Finally, to give a separation between ϵ-DP and ρ-CDP for GEP, we also provide the lower bound $\Omega(\frac{d}{n} + \frac{d^2}{n^2\epsilon^2})$ and $\Omega(\frac{s\log d}{n} + \frac{s^2\log^2 d}{n^2\epsilon^2})$ of private minimax risk for PCA, under the statistical setting and ϵ-DP model, in low and high dimensional sparse case respectively. [1]

[1]The first two authors contributed equally.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 INTRODUCTION

(Sparse) generalized eigenvalue problem (GEP) receives much attention recently as it arises in a number of standard and modern statistical learning models, including (sparse) principal component analysis (PCA), (sparse) Fisher's discriminant analysis (FDA), and (sparse) canonical correlation analysis (CCA), which have enormous applications in biomedicine [Liu and Altman, 2015], biomedical imaging [Strickert et al., 2009] and genomics [Parkhomenko et al., 2009].

The wide applications of GEP also present some new challenges to this problem. Particularly, due to the existence of sensitive data (such as biomedical images) and their distributed nature in many applications like biomedicine and genomics, it is often challenging to preserve the privacy of such data as they are extremely difficult to aggregate and learn from. One promising direction is to use some differentially private mechanisms to conduct the aggregation and learning tasks. Differential Privacy (DP) [Dwork et al., 2006] is a commonly-accepted criterion that provides provable protection against identification and is resilient to arbitrary auxiliary information that might be available to attackers. To design DP algorithms, previous work always focus on specific statistical models, such as (sparse) PCA, CCA. However, there is no general framework which can solve them all together. As the above problems all can be formulated as a GEP problem, a DP algorithm for (sparse) GEP could simultaneously solve PCA, CCA, FDA etc. However, to our best knowledge, there is no work on the designing DP algorithms for (sparse) GEP and the theoretical behaviors of GEP in DP model is still unknown.

To address the above issues, in this paper, we provide a first study of GEP under the DP constraint, *i.e., DP-GEP*, under both low dimension and high dimensional sparse settings. Specifically, our contributions can be summarized as following.

- We first consider DP-GEP in the low dimensional case. Specifically, we propose a ρ-Concentrated DP (CDP) method, namely DP-Rayleigh Flow, and show that if the initial vector is close enough the optimal one, then the output of algorithm could achieve an

ℓ_2-norm estimation error of $\tilde{O}(\frac{d}{n^2\rho})$ and $\tilde{O}(\frac{d}{n}+\frac{d}{n^2\rho})$ in the deterministic and statistical setting respectively (under some mild assumptions), where n is the sample size and d is the dimension of the space. Moreover, we also show that if n is sufficiently large, then we can efficiently find such an initial parameter with ρ-CDP guarantee by reformulating the original GEP problem as a convex programming problem with a LASSO penalty.

- We then consider the problem in the high dimensional case with $d \gg n$, where we assume the underlying parameter is s-sparse. Particularly, we present a method namely DP-Truncated Rayleigh Flow which could achieve an error of $\tilde{O}(\frac{s\log d}{n^2\rho})$ and $\tilde{O}(\frac{s\log d}{n}+\frac{s\log d}{n^2\rho})$ in deterministic and statistical setting respectively, with some initial parameter. As corollaries, we also provide the first theoretical result for CCA, FDA and Sliced Inverse Regression (SIR) in the CDP model.

- We also study the lower bounds of DP-GEP under various settings. We first show that the previous upper bounds in the stochastic setting are optimal up to a factor of $\text{Poly}(\log n)$ by showing the optimal rates of private minimax risk for PCA and SIR in the CDP model. Then we study the ϵ-DP model and show that the private minimax risk for ϵ-DP-PCA is lower bounded by $\Omega(\frac{d}{n}+\frac{d^2}{n^2\epsilon^2})$ and $\Omega(\frac{s\log d}{n}+\frac{s^2\log^2 d}{n^2\epsilon^2})$ in low dimensional and high dimensional setting respectively. Compared with our upper bounds, we can see a separation of the problem in ϵ-DP and CDP. To the best of our knowledge, these are first lower bounds of DP sparse PCA and DP-SIR under statistical setting, which may could used to other problems. Finally, extensive experiments on both synthetic and real-world data support our previous theoretical analysis.

Due to space limit, the full version of some theorems, all proofs and experiments are included in Appendix.

2 RELATED WORK

As we mentioned earlier, there is no previous result on DP-GEP, and there is even no result on DP-FDA and DP-SIR. For DP-CCA, [Imtiaz and Sarwate, 2017] first studies the problem, which is later extended by [Imtiaz and Sarwate, 2019, Shen, 2020] to other settings. However, their algorithms cannot be extended to the high dimensional sparse case and there is no theoretical guarantees for their methods. Below we will focus on the previous results on DP-PCA.

There is a vast number of papers studying PCA under differential privacy, starting from the SULQ framework [Blum et al., 2005, Dwork et al., 2014, Chaudhuri et al., 2013, Gonem and Gilad-Bachrach, 2018, Ge et al., 2018, Balcan et al., 2016]. For DP-PCA in (ϵ,δ)-DP model, [Hardt and Roth, 2013, Balcan et al., 2016, Hardt and Price, 2014] study noisy versions of the power method. [Dwork et al., 2014] considers the deterministic setting and provides the optimal rate of the problem for general K-PCA. However, all these methods cannot be extended to the high dimensional sparse case. For high dimensional sparse PCA, [Ge et al., 2018] studies the problem in the distributed setting and proposed a noisy iterative hard thresholding power method, and [Wang and Xu, 2020] focuses on the problem in the local DP model by showing its upper bound and lower bound. However, these methods are only for PCA and cannot be extended to GEP where here we have an additional constraint which also depends on the dataset. Moreover, the proof of lower bound is also different since it only focuses on the local DP model while in this paper we study the central one.

There are also several papers provide the lower bounds of PCA in central ϵ-DP model. However, all of them are different with ours. Specifically, [Chaudhuri et al., 2012] studies the deterministic setting and shows the lower bound of $\Omega(\frac{d^2}{n^2\epsilon^2})$ for the estimation error, which is later extended by [Kapralov and Talwar, 2013] to general K-PCA case. Compared with their results, we consider the stochastic setting instead and show the lower bound of $\Omega(\frac{d}{n}+\frac{d^2}{n^2\epsilon^2})$. Due to different settings, their proof techniques cannot be used to ours and we use a different technique of proof. [Liu et al., 2022] recently also studies the lower bound of ϵ-DP-PCA under statistical setting. However, their assumptions on the underlying distribution of data are totally different with ours, which indicates that our results are incomparable with theirs. Moreover, they only consider the low dimensional case while we consider both low dimension and high dimensional sparse cases. For PCA in the central (ϵ,δ)-DP model, [Dwork et al., 2014] provides a lower bound of $\Omega(\frac{d\log\frac{1}{\delta}}{n^2\epsilon^2})$ for the problem in the deterministic setting by using the fingerprinting codes while in this paper we provide the lower bound of $\Omega(\frac{d}{n}+\frac{d}{n^2\rho})$ under the stochastic setting and in the CDP model. We also consider the high dimensional sparse case.

3 PRELIMINARIES

Notations: We denote $\lambda_i(Z), \lambda_{\max}(Z), \lambda_{\min}(Z)$ as the i-th, maximal and minimal eigenvalue of matrix Z respectively. And denote the condition number of a positive definite matrix $Z \in \mathbb{R}^{d\times d}$ as $\kappa(Z) = \frac{\lambda_{\max}(Z)}{\lambda_{\min}(Z)}$. Moreover, let λ_j and $\hat{\lambda}_j$ be the j-th generalized eigenvalue of the matrix pairs (A,B) and $(\hat{A},\hat{B})$ respectively. Given an index set $F \subseteq [d]$, let $Z_F \in \mathbb{R}^{|F|\times|F|}$ be the submatrix of Z where the rows and columns are restricted to the set F. We also denote $\rho(Z,s) = \sup_{\|u\|_2=1,\|u\|_0\le s}|u^TZu|$ and $\rho(Z) = \|Z\|_2 = \rho(Z,d)$. For a pair of symmetric ma-

trix (A, B) we denote its Crawford number as $\text{cr}(A, B) = \min_{v:||v||_2=1} \sqrt{(v^T Av)^2 + (v^T Bv)^2} \geq 0$.

In this section, we recall some definitions related to Differnetial Privacy and Generalized Eigenvalue Problem.

Definition 1 (Differential Privacy [Dwork et al., 2006])**.** Given a data universe $\mathcal{X}$, we say that two datasets $D, D' \subseteq \mathcal{X}$ are neighbors if they differ by only one data sample, which is denoted as $D \sim D'$. A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all neighboring datasets D, D' and for all events S in the output space of $\mathcal{A}$, we have $\Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon}\Pr(\mathcal{A}(D') \in S) + \delta$. When $\delta = 0$ we call the algorithm is ϵ-DP.

Definition 2 (Concentrated DP [Bun and Steinke, 2016])**.** A randomized algorithm $\mathcal{A}$ is ρ-Concentrated DP (CDP) if for all neighboring datasets D, D' and for all $\alpha > 1$ we have $D_\alpha(\mathcal{A}(D)\|\mathcal{A}(D')) \leq \alpha\rho$, where $D_\alpha(\mathcal{A}(D)\|\mathcal{A}(D'))$ is the α-Rényi divergence between $\mathcal{A}(D)$ and $\mathcal{A}(D')$.

Actually, CDP lives between ϵ-DP and (ϵ, δ)-DP:

Lemma 1 ([Bun and Steinke, 2016])**.** For every $\epsilon > 0$, if algorithm $\mathcal{A}$ is ϵ-DP then it will be $\frac{\epsilon^2}{2}$-CDP. If $\mathcal{A}$ is ρ-CDP, then it will be (ϵ, δ)-DP with $\epsilon = \rho + 2\sqrt{\rho \log \frac{1}{\delta}}$.

By the previous lemma, we can see to achieve a given (ϵ, δ)-DP guarantee, it is sufficient to show the algorithm is $\rho = (\sqrt{\epsilon + \log 1/\delta} - \sqrt{\log 1/\delta})^2 \approx \frac{\epsilon^2}{4 \log 1/\delta}$-CDP. **Thus, all ρ-CDP algorithms with their utility in this paper can be transformed to the (ϵ, δ)-DP version with $\log 1/\delta \gg \epsilon$ by simply replacing ρ by $\frac{\epsilon^2}{4 \log 1/\delta}$.**

In this paper, we will mainly use the Gaussian mechanism and the Composition Theorem to guarantee CDP.

Definition 3 (Gaussian Mechanism)**.** Given any function $q : \mathcal{X}^n \to \mathbb{R}^d$, the Gaussian mechanism is defined as $q(D) + \xi$ where $\xi \sim \mathcal{N}(0, \frac{\Delta_2^2}{2\rho}\mathbb{I}_d)$, where where $\Delta_2(q)$ is the ℓ_2-sensitivity of the function q, *i.e.,* $\Delta_2(q) = \sup_{D \sim d'} ||q(D) - q(D')||_2$. Gaussian mechanism preserves ρ-CDP.

Lemma 2 (Composition Theorem)**.** If $\mathcal{A}$ is an adaptive composition of CDP algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_T$ where $\mathcal{A}_i$ is ρ_i-CDP. Then $\mathcal{A}$ will be ρ-CDP with $\rho = \sum_{i=1}^T \rho_i$.

Definition 4 (GEP [Golub and Van Loan, 1996])**.** The generalized eigenvalues of the symmetric-definite pair (A, B) are denote by $\lambda(A, B) = \{\lambda | \det(A - \lambda B) = 0\}$. If $\lambda \in \lambda(A, B)$ and v is a non-zero vector satisfies $Av = \lambda Bv$, then v is a generalized eigenvector.

Given an n-size data set $D = \{x_1, \cdots, x_n\}$, matrices $\hat{A}$ and $\hat{B}$. The (largest) generalized eigenvalue problem (GEP) of $(\hat{A}, \hat{B})$ is characterized as

$$\tilde{v} = \arg\max_{v \in \mathbb{R}^d} v^T \hat{A} v, \text{ s.t. } v^T \hat{B} v = 1, \tag{1}$$

where $\hat{A} = \hat{A}(D) \in \mathbb{R}^{d \times d}$ and $\hat{B} = \hat{B}(D) \in \mathbb{R}^{d \times d}$ are matrices that (may) dependent on the dataset D. Besides the **deterministic setting**, for some statistical models we also want to study the **stochastic setting** where we assume each record is sampled from some underlying unknown distribution $\mathcal{P}$. And our goal is to solve the following problem based on the data D, where $A = \mathbb{E}[\hat{A}]$ and $B = \mathbb{E}[\hat{B}]$.

$$v^* = \arg\max_{v \in \mathbb{R}^d} v^T A v, \text{ s.t. } v^T B v = 1. \tag{2}$$

In the high dimensional setting, we assume $d \gg n$ and the underlying parameter v^* in (2) or $\tilde{v}$ in (1) has an additional sparsity structure, *i.e.,* we assume $\|v^*\|_0 \leq s$ or $\|\tilde{v}\|_0 \leq s$ for some $s \ll d$. Now the sparse GEP becomes to

$$\tilde{v}_s = \arg\max_{v \in \mathbb{R}^d} v^T \hat{A} v, \text{ s.t. } v^T \hat{B} v = 1, \|v\|_0 \leq s. \tag{3}$$

$$v_s^* = \arg\max_{v \in \mathbb{R}^d} v^T A v, \text{ s.t. } v^T B v = 1, \|v\|_0 \leq s. \tag{4}$$

In the following, we will provide some statistical models that are special cases of (sparse) GEP.

Principal Component Analysis (PCA): Given dataset $D = \{x_1, \cdots, x_2\}$ with each $x_i \in \mathbb{R}^d$, (sparse) PCA can be formulated as (sparse) GEP with $\hat{B} = I_d$ and $\hat{A} = \hat{\Sigma}$ where $\hat{\Sigma}$ is the covariance matrix $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T$ with $\mu = \frac{\sum_{i=1}^n x_i}{n}$. In the stochastic setting A is the polulation version of $\hat{A}$, *i.e.,* $A = \mathbb{E}[(x-\mu)(x-\mu)^T]$ with $\mu = \mathbb{E}[x]$.

Canonical Component Analysis (CCA): Given dataset $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ with each $x_i \in \mathbb{R}^{d_1}$ and $y_i \in \mathbb{R}^{d_2}$, (sparse) CCA can be formulated as (sparse) GEP with

$$\hat{A} = \begin{pmatrix} 0 & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{xy} & 0 \end{pmatrix}, \hat{B} = \begin{pmatrix} \hat{\Sigma}_x & 0 \\ 0 & \hat{\Sigma}_y \end{pmatrix},$$

where $\hat{\Sigma}_x = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_x)(x_i - \mu_x)^T$, $\hat{\Sigma}_y = \frac{1}{n}\sum_{i=1}^n (y_i - \mu_y)(y_i - \mu_y)^T$ and $\hat{\Sigma}_{xy} = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y)^T$ with $\mu_x = \frac{\sum_{i=1}^n x_i}{n}$ and $\mu_y = \frac{\sum_{i=1}^n y_i}{n}$. In the stochastic setting, A and B are the population version of $\hat{A}$ and $\hat{B}$ respectively.

Fisher's Discriminant Analysis (FDA): Given n samples with K different classes, Fisher's discriminant analysis seeks a low dimensional projection of the samples such that the between-class variance is large relative to the with-class variance. Specifically, it could be formulated as GEP with

$$\hat{A} = \frac{1}{n}\sum_{k=1}^K \sum_{i \in C_k} (x_i - \hat{u}_k)(x_i - \hat{u}_k)^T,$$
$$\hat{B} = \frac{1}{n}\sum_{k=1}^K n_k \hat{u}_k \hat{u}_k^T, \hat{u}_k = \sum_{i \in C_k} \frac{x_i}{n_k}, \tag{5}$$

where C_k is the index set for the k-th class, *i.e.,* if $i \in C_k$ then x_i is in the k-th class, and $n_k = |C_k|$.

Sliced Inverse Regression (SIR): In SIR we have the statistical model $Y = f(v_1^T X, \cdots, v_k^T X, \zeta)$, where ζ is some random noise and is independent on X, $f(\cdot)$ is some unknown link function. It has been shown that under some mild assumptions, the space that is spanned by $v_1, \cdots, v_k$ can be identified [Li, 1991]. Particularly, the first leading eigenvector of the subspace that is spanned by $v_1, \cdots, v_k$ can be identified by solving the GEP with A be the covariance matrix of the conditional expectation $\mathbb{E}(X|Y)$ and B as the covariance matrix of X. That is:

$$\begin{aligned}
&\hat{A} = \hat{\Sigma}_{E(X|Y)}, \hat{B} = \hat{\Sigma}_x, \hat{\Sigma}_{E(X|Y)} = \hat{\Sigma}_x - E[\hat{\Sigma}_{(x|y)}] \\
&\hat{\Sigma}_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^T, \\
&\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i, u_k = \sum_{i \in C_k} \frac{x_i}{n_k} \\
&E[\hat{\Sigma}_{(x|y)}] = \frac{1}{n}\sum_{k=1}^{K}\sum_{x \in C_k} n_k \frac{(x_i - \mu_k)(x_i - \mu_k)^T}{n_k}
\end{aligned} \tag{6}$$

In the following we present the definition of DP-GEP.

Definition 5 (DP-GEP)**.** Given a dataset $D = \{x_1, \cdots, x_n\}$ and its corresponding (sparse) GEP, the goal of Differentially Private GEP (GEP) is finding a private estimator v_{priv} based on some DP algorithm. Moreover, we want our private estimator close enough to the optimal parameter. Specifically, in this paper, we will mainly use the similarity $1 - \frac{\langle v_{\text{priv}}, v^* \rangle}{\|v_{\text{priv}}\|_2 \|v^*\|_2}$ to the measure the closeness. Based on different settings, v^* could be the optimal vector of the problem (1), (2), (3) or (4).

If we denote $\mathbf{E_A} = \mathbf{A} - \hat{\mathbf{A}}$ and $\mathbf{E_B} = \mathbf{B} - \hat{\mathbf{B}}$. Then we can see that the deterministic setting is a special case of the stochastic setting with $E_A = E_B = 0$. Thus, in this paper we mainly focus on the stochastic setting. Next we propose several assumptions that will be used throughout the paper. Assumption 1 requires that the Frobenius norm sensitivity of $\hat{A}$ and $\hat{B}$ are bounded by $O(\frac{1}{n})$.

Assumption 1. Given any neighboring datasets D and D'. For $\hat{A}(D) \in \mathbb{R}^{d \times d}$ and $\hat{B}(D) \in \mathbb{R}^{d \times d}$ in problem (1) we assume $\|\hat{A}(D) - \hat{A}(D')\|_F \leq \frac{C_1}{n}$ and $\|\hat{B}(D) - \hat{B}(D')\|_F \leq \frac{C_2}{n}$ for some constants $C_1, C_2 \geq 0$.

The following assumption is to control the norm of the error matrix E_A and E_B in the statistical setting.

Assumption 2. We assume that for any $0 < s \leq d$ we have $\rho(E_A, s), \rho(E_B, s) = O(\sqrt{\frac{s \log d}{n}})$.

It is notable that Assumption 2 is only used in the utility analysis throughout the paper and is only for simplicity, i.e., the privacy guarantees will still hold even Assumption 2 does not hold. Moreover, all of our utility analysis could be extended to general $\rho(E_A), \rho(E_B), \rho(E_A, s), \rho(E_B, s)$, see Appendix for details. In the following we show that the above assumptions hold (with high probability) for all the statistical models we mentioned previously if each data sample $\|x_i\|_2 \leq 1$. Thus, we can see these two assumptions are mild.

Theorem 1. If each $\|x_i\|_2 \leq 1$ for $i \in [n]$, then PCA, CCA, FDA and SIR all satisfy Assumption 1. Moreover, PCA, CCA and SIR satisfy Assumption 2 (with high probability).

4 LOW DIMENSION CASE

In this section we consider the low dimension case, *i.e.,* problem (1) and (2). To illustrate our idea, we first review the classical method for GEP by using Rayleigh's quotient [Parlett, 1998]. Specifically, problem (1) can be rewritten as

$$\max_{v \in \mathbb{R}^d} J(v) = \frac{v^T \hat{A} v}{v^T \hat{B} v}, \tag{7}$$

where the objective function could be seen as the generalized Rayleigh quotient. To solve (7), one can use the Gradient Ascent method, i.e., in the t-th iteration the vector v_t is updated as

$$v_t = v_{t-1} + \eta \nabla_v J(v_{t-1}),$$

where $\nabla_v J_v(v_{t-1}) \propto \hat{A}v_{t-1} - \frac{v_{t-1}^T \hat{A} v_{t-1}}{v_{t-1}^T \hat{B} v_{t-1}} \hat{B} v_{t-1}$ and η is the stepsize. Thus, to design DP methods, one natural approach is based on the idea of DP-SGD, which is a commonly used method for DP Empirical Risk Minimization (ERM) and DP Deep Learning [Abadi et al., 2016, Wu et al., 2017, Wang et al., 2018, Bassily et al., 2014]. The idea of DP-SGD is injecting some Gaussian noise into the (stochastic) gradient in each iteration. That is

$$v_t = v_{t-1} + \eta(\nabla_v J(v_{t-1}) + \zeta_{t-1}),$$

where ζ_{t-1} is a Gaussian vector where the variance of each coordinate is proportional to the sensitivity of $\nabla_v J(v_{t-1})$. However, the main challenge is that, unlike the objective functions in ERM or Deep Learning where the sensitivity of gradient is $O(\frac{1}{n})$, our objective function (Rayleigh quotient) cannot be decomposed into a sum of loss functions, which means the sensitivity of $\nabla_v J(v_{t-1})$ is larger and even could be unbounded. Thus, we cannot use DP-SGD based methods and need new approaches.

Based on the specific structure of $\nabla_v J_v(v_{t-1})$, here we propose a new method namely DP-Rayleigh Flow. Specifically, instead of injecting noise to the gradient, we add noises to matrices $\hat{A}$ and $\hat{B}$ in each iteration, see Algorithm 1 for details. However, compared with the original Rayleigh Flow method as mentioned above, we need

some modifications. First, instead of using some fixed stepsize η, in each iteration of Algorithm 1 we rescale it by $\rho_t = v_{t-1}^T \tilde{A}^t v_{t-1} / v_{t-1}^T \tilde{B}^t v_{t-1}$, where $\tilde{A}^t$ and $\tilde{B}^t$ are perturbed matrices in the t-th iteration, *i.e.,* we use $\frac{\eta}{\rho_t}$ as the stepsize, which is convenient for our following theoretical analysis. Secondly, after updating by using Gradient Ascent *i.e.,* calculating $C^t v_{t-1}$, in step 5 of Algorithm 1 we need to normalize the vector to ensure v_t has unit ℓ_2-norm. This step guarantees that the generalized Rayleigh quotient for the updated vector is at least as large as that of the initial vector. In the following we provide theoretical guarantees for our algorithm.

Algorithm 1 DP-Rayleigh Flow

1: **Input:** Matrices $\hat{A}$ and $\hat{B}$, initial parameter v_0 with $\|v_0\|_2 = 1$, step size η (will be specified later), iteration numbers m, privacy parameter ρ.
2: **for** $t = 1, \cdots, m$ **do**.
3: Denote $\tilde{A}^t = \hat{A} + Z_1^t$, $\tilde{B}^t = \hat{B} + Z_2^t$, where Z_1^t and Z_2^t are symmetric matrix where the upper triangle (including the diagonal) is i.i.d. samples from $\mathcal{N}(0, \sigma_1^2)$ with $\sigma_1^2 = \frac{C_1^2 m}{n^2 \rho}$ and $\mathcal{N}(0, \sigma_2^2)$ with $\sigma_2^2 = \frac{C_2^2 m}{n^2 \rho}$ respectively, and each lower triangle entry is copied from its upper triangle counterpart.
4: Denote $\rho_t = \frac{v_{t-1}^T \tilde{A}^t v_{t-1}}{v_{t-1}^T \tilde{B}^t v_{t-1}}$ and $C^t = I + \frac{\eta}{\rho_t}(\tilde{A}^t - \rho_t \tilde{B}^t)$
5: Update $v_t = \frac{C^t v_{t-1}}{||C^t v_{t-1}||_2}$.
6: **end for**
7: **return** v_m.

Theorem 2. Under Assumption 1, for any $\rho > 0$ Algorithm 1 is ρ-CDP.

Before showing the estimation error of the output in Algorithm 1, we first introduce several notations and assumptions. The following theorem indicates that when n is sufficiently large, the generalized eigenvalue of the perturbed matrices is close to the generalized eigenvalue of the underlying matrices.

Theorem 3. Let $\tilde{\lambda}_k^t$ be the kth generalized eigenvalues of $(\tilde{A}^t, \tilde{B}^t)$, where $(\tilde{A}^t, \tilde{B}^t)$ are the perturbed matrices in the t-th iteration. Under Assumption 2, given any failure probability $\zeta > 0$, let constants $0 \le b < \min_{j \in [d]} \frac{\lambda_j}{2\lambda_j^2 + 1}$, $0 \le c$ and if n is sufficiently large such that, $n \ge \tilde{\Omega}(\max\{\frac{d}{c^2 \lambda_{\min}^2(B)}, \frac{d}{b^2 \mathrm{cr}^2(A,B)}, \frac{\sqrt{dm \log \frac{1}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{dm \log \frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$. Then with probability at least $1 - \zeta$, there exists constants a such that for all $t \in [m]$,

$$(1-a)\lambda_j \le \tilde{\lambda}_j^t \le (1+a)\lambda_j,$$
$$(1-c)\lambda_j(B) \le \lambda_j(\tilde{B}^t) \le (1+c)\lambda_j(B) \quad (8)$$
$$C_{\text{lower}}\kappa(B) \le \kappa(\tilde{B}^t) \le C_{\text{upper}}\kappa(B) \quad (9)$$

where $C_{\text{lower}} = \frac{1-c}{1+c}$, $C_{\text{upper}} = \frac{1+c}{1-c}$ Furthermore, we have $\tilde{\lambda}_2^t \le \gamma \tilde{\lambda}_1^t$, where $\gamma = \frac{(1+a)\lambda_2}{(1-a)\lambda_1}$.

Theorem 4 (Informal)**.** Under Theorem 3 and choose the stepsize η such that $\eta \lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B) + \gamma}} < \frac{1}{2}.$$

Then if n is sufficient large, in Algorithm 3 we set $m = O(\log n)$ and if the input vector v_0 with $\|v_0\|_2 = 1$ satisfying $\frac{|\langle v^*, v_0\rangle|}{\|v^*\|_2} \ge 1 - \frac{\theta(A,B)}{2}$ with

$$\theta(A,B) = \min\{\frac{1}{8C_{\text{upper}}\kappa(B)}, \frac{1/\gamma - 1}{3C_{\text{upper}}\kappa(B)}, \frac{1-\gamma}{30(1+c)C_{\text{upper}}^2\eta\lambda_{\max}(B)\kappa^2(B)\{C_{\text{upper}}\kappa(B)+\gamma\}}\}, \quad (10)$$

we have the following with probability at least $1 - \zeta$,

$$1 - \frac{\langle v^*, v_m\rangle}{\|v^*\|_2} \le O(\frac{\theta(A,B)}{(1-\nu)^2} \times \big(\frac{1}{\lambda_{gap}^2 \mathrm{cr}^2(A,B)}\frac{d\log d}{n} + \frac{1}{\lambda_{gap}^2 \mathrm{cr}^2(A,B)}\frac{d\log n \log d \log\frac{1}{\zeta}}{n^2\rho}\big)), \quad (11)$$

where

$$\lambda_{gap} = \min_{j>1} \frac{\lambda_1 - (1+a)\lambda_j}{\sqrt{1+\lambda_1^2}\sqrt{1+(1-a)^2\lambda_j^2}} \quad (12)$$

is the eigengap for the GEP.

Similarly, for the deterministic setting where $\hat{A} = A, \hat{B} = B$, if n is sufficiently large and we set some appropriate parameters in Algorithm 3, with probability at least $1 - \zeta$

$$1 - \frac{\langle \tilde{v}, v_m\rangle}{\|\tilde{v}\|_2} \le \tilde{O}(\frac{\theta(A,B)}{(1-\nu)^2\lambda_{gap}^2 \mathrm{cr}^2(A,B)}\frac{d\log d\log\frac{1}{\zeta}}{n^2\rho}). \quad (13)$$

Remark 1. Since $\theta(A,B), \lambda_{gap}, \mathrm{cr}(A,B)$ and v all only depend on the underlying matrices A and B. Thus the output could achieve an error of $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ and $\tilde{O}(\frac{d}{n^2\rho})$ under the stochastic setting and deterministic setting respectively (if we omit other terms). Note that in the non-private case, the optimal rate is $O(\frac{d}{n})$ for many statistical models such as PCA or CCA [Cai et al., 2013, Gao et al., 2015] if each $\|x_i\|_2 \le O(1)$. Thus, based on Theorem 1 we can see it is possible to obtain privacy nearly for free when $\rho > \frac{1}{n}$ in the statistical setting.

One major issue in Theorem 4 is we need to assume the initial vector v_0 is close enough to v^* such that $\frac{|\langle v^*, v_0\rangle|}{\|v^*\|_2} \ge 1 - \frac{\theta(A,B)}{2}$. In general, this condition is necessary since in general GEP is non-concave and the Gradient Ascent

method can only ensure the parameter converges to some local maximum. However, with some additional assumptions and n is the sufficiently large, in the following we show how to find such an initial vector privately and efficiently.

Note that in the non-private case, originally finding the K leading generalized eigenvectors for matrix pair $(\hat{A}, \hat{B})$ is equivalent to solve the following optimization problem:

$$\min_{U\in\mathbb{R}^{d\times K}} -\text{tr}(U^T\hat{A}U), \text{ s.t. } U^T\hat{B}U = I_K. \quad (14)$$

Due to the non-convexity of the previous problem, motivated by [Tan et al., 2018, Vu et al., 2013a, Wang and Xu, 2020] here we consider a convex relaxation with a LASSO penalty, i.e.,

$$\min_{P\in\mathbb{R}^{d\times d}} -\text{tr}(\hat{A}P) + \phi\|P\|_{1,1},$$
$$\text{s.t. } \|\hat{B}^{\frac{1}{2}}P\hat{B}^{\frac{1}{2}}\|_{nu} \le K, \|\hat{B}^{\frac{1}{2}}P\hat{B}^{\frac{1}{2}}\|_2 \le 1, \quad (15)$$

where for a matrix A, $\|A\|_{nu}$ is defined as the sum of its singular values, $A^{\frac{1}{2}}$ is the square root of A, and $\|A\|_{1,1}$ is the ℓ_1-norm of the vector of row-wise ℓ_1 norm of A.

Our private estimator is based on (15). That is, instead of using the empirical matrices $\hat{A}$ and $\hat{B}$, we use their perturbed version to ensure DP. Specifically, we will solve the following optimization problem:

$$\hat{P} = \arg\min_{P\in\mathbb{R}^{d\times d}} -\text{tr}(\tilde{A}P) + \phi\|P\|_{1,1},$$
$$\text{s.t. } \|\tilde{B}^{\frac{1}{2}}P\tilde{B}^{\frac{1}{2}}\|_{nu} \le K, \|\tilde{B}^{\frac{1}{2}}P\tilde{B}^{\frac{1}{2}}\|_2 \le 1, \quad (16)$$

where $\tilde{A} = \hat{A} + Z_1$, $\tilde{B} = \hat{B} + Z_2$ and Z_1 and Z_2 are symmetric Gaussian matrices to ensure DP.

Since the optimization problem (16) is convex, we can follow the approach in [Wang and Xu, 2020] to solve it by using ADMM method (see Algorithm 2 for the details).

Informally we have the following result.

Theorem 5 (Informal)**.** Under Assumption 1, the solution of the optimization problem (16) is ρ-CDP. Moreover, under Assumption 2 and assume that $\|E_A\|_{\infty,\infty}, \|E_B\|_{\infty,\infty} = O(\sqrt{\frac{\log d}{n}})$, and n is sufficiently large, take $\phi = \tilde{O}(\lambda_{\max}(B)\lambda_1(\frac{\sqrt{d}}{\sqrt{n}} + \frac{\sqrt{d}}{n\sqrt{\rho}}))$, $K = 1$ in (16). Then the largest eigenvalue of the matrix $\hat{P}$ which is denoted as v_0, satisfies $\langle v_0, v^*\rangle \ge 1 - \theta(A,B)/2$ with high probability. For a matrix A, $\|A\|_{\infty,\infty}$ is defined as the maximal absolute value among the entries in A.

In the above theorem we need to assume that $\|E_A\|_{\infty,\infty}$, $\|E_B\|_{\infty,\infty} = O(\sqrt{\frac{\log d}{n}})$, these assumptions hold in the deterministic setting where $E_A = E_B = 0$. In the stochastic setting, we can show these assumptions hold for PCA, CCA and SIR if $\|x_i\|_2 \le 1$ (see the Proof of Theorem 1).

Algorithm 2 Privately Finding an Initial Vector

1: **Input:** Matrices $\hat{A}$ and $\hat{B}$, privacy parameters ρ, tuning parameter ϕ, ADMM parameter v, and convergence criterion β.
2: Initialize matrices P_0, H_0 and Γ_0. Set $t = 0$
3: Let $\tilde{A} = \hat{A} + Z_1$, $\tilde{B} = \hat{B} + Z_2$ and Z_1 and Z_2 are symmetric matrix where the upper triangle (including the diagonal) is i.i.d. samples from $\mathcal{N}(0, \sigma_1^2)$ with $\sigma_1^2 = \frac{C_1^2}{2n^2\rho}$ and $\mathcal{N}(0, \sigma_2^2)$ with $\sigma_2^2 = \frac{C_2^2}{2n^2\rho}$ respectively, and each lower triangle entry is copied from its upper triangle counterpart.
4: Update P by solving the following lasso problem:

$$P_{t+1} = \arg\min \frac{v}{2}\|\tilde{B}^{\frac{1}{2}}P\tilde{B}^{\frac{1}{2}} - H_t + \Gamma_t\|_F^2 - \text{tr}(\tilde{A}P) + \phi\|P\|_{1,1}.$$

5: Let $\sum_{i=1}^d w_j a_j a_j^T$ be the singular value decomposition of $\Gamma_t + \tilde{B}^{\frac{1}{2}}P_{t+1}\tilde{B}^{\frac{1}{2}}$ and let

$$\gamma^* = \arg\min_{\gamma>0} \gamma,$$
$$\text{s.t. } \sum_{j=1}^d \min\{1, \max\{w_j - \gamma, 0\}\} \le K.$$

Update H by $H_{t+1} = \sum_{j=1}^d \min\{1, \max\{w_j - \gamma^*, 0\}a_j a_j^T$.
6: Update Γ as $\Gamma_{t+1} = \Gamma_t + \tilde{B}^{\frac{1}{2}}P_{t+1}\tilde{B}^{\frac{1}{2}} - H_{t+1}$.
7: If $\|P_{t+1} - P_t\|_F > \beta$, let $t = t + 1$ and repeat the procedure 4-6.
8: **return** The leading eigenvector of P_{t+1}.

5 HIGH DIMENSIONAL SPARSE CASE

In the previous section, we showed the upper bounds of the estimation error in stochastic and deterministic settings. However, in the high dimensional case where $d \gg n$, the previous two bounds will be quite large so that their rates become trivial. To address the high dimensionality issue, in this section we consider the sparse GEP instead, *i.e.,* problem (3) and (4). Specifically, we propose a truncated version of Algorithm 1, namely DP-Truncated Rayleigh Flow, see Algorithm 3 for details. Compared with Algorithm 1, there is an additional truncation step. That is, we select the indices with largest k magnitude of the vector, keep the entries of vectors among these indices and let the remain entries be zero. Intuitively, the truncation step could project the vector onto a low dimensional space (and thus the effective dimension now becomes to k instead of d), and it will diminish the noises we added to $\hat{A}$ and $\hat{B}$. Note that the idea of truncating the vector to enforce it be sparse has also been used in other DP machine learning problems, such as [Cai et al., 2019,

Algorithm 3 DP-Truncated Rayleigh Flow

Input: Matrices $\hat{A}$ and $\hat{B}$, sparsity k, initial parameter v_0 is a k-sparse vector with $\|v_0\|_2 = 1$, step size η, iteration number m, privacy parameter ρ.
for $t = 1, \cdots, m$ **do**.
  Denote $\tilde{A}^t = \hat{A} + Z_1^t$, $\tilde{B}^t = \hat{B} + Z_2^t$, where Z_1^t and Z_2^t are symmetric matrix where the upper triangle (including the diagonal) is i.i.d. samples from $\mathcal{N}(0, \sigma_1^2)$ with $\sigma_1^2 = \frac{C_1^2 m}{n^2\rho}$ and $\mathcal{N}(0, \sigma_2^2)$ with $\sigma_2^2 = \frac{C_2^2 m}{n^2\rho}$ respectively, and each lower triangle entry is copied from its upper triangle counterpart.
  Denote $\rho_t = \frac{v_{t-1}^T \tilde{A}^t v_{t-1}}{v_{t-1}^T \tilde{B}^t v_{t-1}}$ and $C^t = I + (\eta/\rho_t)(\tilde{A}^t - \rho_t \tilde{B}^t)$.
  Update $v_t' = \frac{C^t v_{t-1}}{||C^t v_{t-1}||_2}$.
  Let $F_t = \text{supp}(v_t', k)$ be the set of indices of v_t' with the largest k absolute values.
  Denote $\hat{v}_t = \text{truncate}(v_t', F_t)$, *i.e.,* $\hat{v}_t$ is the truncated vector of v_t' by setting $(v_t')_i = 0$ if $i \notin F_t$.
  Update $v_t = \frac{\hat{v}_t}{\|\hat{v}_t\|_2}$.
end for
return v_m.

Wang et al., 2019, Wang and Gu, 2020, Hu et al., 2021] for DP-ERM and [Ge et al., 2018] for DP-Sparse PCA. However, as we mentioned, unlike those objective functions, the Rayleigh quotient cannot be decomposed as a sum of functions, it is unknown whether truncation step is indeed helpful. We will provide an affirmative answer in this section.

Theorem 6. Under Assumption 1, for any $0 < \rho$, Algorithm 3 is ρ-CDP.

Before providing the estimation error of Algorithm 3 we first provide the following notations.

Notations: For v_s^* in (4) we denote $V = \text{supp}(v_s^*)$ as the index set corresponding to the non-zero elements of v_s^*. Let $F \subseteq [d]$ be a superset of V with $|F| = k'$, where $k' = 2k + s$ and k is in Algorithm 3. Let $\lambda_j(F)$, $\tilde{\lambda}_k^t(F)$ and $\hat{\lambda}_j(F)$ be the j-th generalized eigenvalue of the matrix pairs (A_F, B_F), $(\tilde{A}_F^t, \tilde{B}_F^t)$ and $(\hat{A}_F, \hat{B}_F)$, respectively. Denote $\text{cr}(k') = \inf_{F:|F|\le k'} \text{cr}(A_F, B_F)$.

Similar to Theorem 3, we first show that when n is sufficiently large, then the generalized eigenvalue (restricted to the set F) of the perturbed matrices is close to the generalized eigenvalue of the underlying matrices.

Theorem 7. Under Assumption 2, given any failure probability $\zeta > 0$, if n is sufficiently large such that $n \ge \Omega(\max\{\frac{k'}{b^2\text{cr}^2(k')}, \frac{k'}{c^2\lambda_{\min}^2(B)}, \frac{\sqrt{k'm\log\frac{1}{\zeta}}}{b\sqrt{\rho}}, \frac{\sqrt{k'm\log\frac{1}{\zeta}}}{c\lambda_{\min}(B)\sqrt{\rho}}\})$ for some constants $c \ge 0$ and $0 \le b < \min_{j\in[d]} \frac{\lambda_j(F)}{2\lambda_j^2(F)+1}$. Then with probability at least $1-\zeta$, there exists constants a and c such that for all $t \in [m]$,

$$(1-a)\lambda_j(F) \le \tilde{\lambda}_j^t(F) \le (1+a)\lambda_j(F), \quad (17)$$

$$(1-c)\lambda_j(B_F) \le \lambda_j(\tilde{B}_F^t) \le (1+c)\lambda_j(B_F), \quad (18)$$

$$C_{\text{lower}}\kappa(B) \le \kappa(\tilde{B}_F^t) \le C_{\text{upper}}\kappa(B) \quad (19)$$

where $C_{\text{lower}} = \frac{1-c}{1+c}$, $C_{\text{upper}} = \frac{1+c}{1-c}$. Furthermore, we have

$$\tilde{\lambda}_2^t(F) \le \gamma\tilde{\lambda}_1^t(F), \quad (20)$$

where $\gamma = \frac{(1+a)\lambda_2(F)}{(1-a)\lambda_1(F)}$.

In the following we provide the statistical error of our private estimator if n is sufficiently large and the initial vector is close the optimal solution with $m = O(\log n)$.

Theorem 8 (Informal)**.** Under Theorem 7 with $k' = 2k+s$ and choose $k = Cs$ for sufficiently large C. In addition, choose stepsize η such that $\eta\lambda_{\max}(B) < \frac{1}{1+c}$ and

$$\nu = \sqrt{1 + 2\sqrt{\frac{s}{k}} + 2\frac{s}{k}} \times \sqrt{1 - \frac{1+c}{8}\eta\lambda_{\min}(B)\frac{1-\gamma}{C_{\text{upper}}\kappa(B)+\gamma}} < \frac{1}{2}.$$

Then if n is sufficiently large, we set $m = O(\log n)$ in Algorithm 3. We have the following with probability at least $1-\zeta$ if the input k-sparse vector v_0 with $\|v_0\|_2$ satisfying $\frac{|\langle v_s^*, v_0\rangle|}{\|v_s^*\|_2} \ge 1 - \frac{\theta(A,B)}{2}$ with $\theta(A,B)$ given in (10).

$$1 - \frac{\langle v_s^*, v_m\rangle}{\|v_s^*\|_2} \le O(\frac{\theta(A,B)}{(1-\nu)^2}(\frac{1}{\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log d}{n} + \frac{1}{\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log n\log d\log\frac{1}{\zeta}}{n^2\rho})). \quad (21)$$

Similarly, for the deterministic setting where $E_A = E_B = 0$ and $\hat{A} = A, \hat{B} = B$, if n is sufficiently large and with some additional mild assumptions. If we set some appropriate parameters in Algorithm 3, with probability at least $1-\zeta$

$$1 - \frac{\langle \tilde{v}_s, v_t\rangle}{\|\tilde{v}_s\|_2} \le \tilde{O}(\frac{\theta(A,B)}{(1-\nu)^2\lambda_{gap}^2\text{cr}^2(k')}\frac{s\log d\log\frac{1}{\zeta}}{n^2\rho}). \quad (22)$$

From Theorem 8 we can find that, the error in the deterministic setting is $\tilde{O}(\frac{s\log d}{n^2\rho})$, while the statistical error of Algorithm 3 will be $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$ (if we omit other terms). These two bounds only depend on logarithmic of d instead of polynomial in the low dimensional case. Moreover, the same as in the low dimensional case, we can obtain privacy for free in the statistical setting.

Corollary 1. If we transform the above upper bounds in CDP to (ϵ, δ)-DP via Lemma 1, we can see for PCA under deterministic setting, the output of Algorithm 1 could

achieve an error of $\tilde{O}(\frac{\sqrt{d\log\frac{1}{\delta}}}{n\epsilon})$, which is near optimal [Dwork et al., 2014]. For sparse PCA under the stochastic setting where A is the covariance matrix and $B = I$, if we further assume that $\|x\|_2 \leq 1$. Then the output of Algorithm 3 could achieve a statistical error of $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d\log\frac{1}{\delta}}{n^2\epsilon^2})$. [Wang et al., 2019] provides the first result on the problem in the local DP model, instead of the central model. Specifically, it shows that the near optimal statistical rate is $\tilde{O}(\frac{s\log d\log\frac{1}{\delta}}{n\epsilon^2})$ under stochastic setting. Compared with our results, we can see a gap between the central and the local model for sparse PCA.

Corollary 2. For the problem of sparse CCA under the stochastic setting and $\|x_2\|_2 \leq 1$. The output of Algorithm 3 could achieve an error of $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$, where $d = d_1 + d_2$. Under deterministic setting, the output of Algorithm 3 could achieve an error of $\tilde{O}(\frac{s\log d}{n^2\rho})$. In the low dimension setting, the error will be $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ and $\tilde{O}(\frac{d}{n^2\rho})$, respectively. Moreover, we have similar results for SIR if $\|x\|_2 \leq 1$. Note that these are the first results on the estimation error for CCA and SIR in the DP model.

Corollary 3. For FDA, the output of Algorithm 1 and 3 could achieve an error of $\tilde{O}(\frac{d}{n^2\rho})$ and $\tilde{O}(\frac{s\log d}{n^2\rho})$ in the low dimension and high dimensional sparse case respectively if $\|x\|_2 \leq 1$. To our best knowledge, this is the first theoretical result for FDA in the DP model.

Similar to the low dimension case, here we still need a good initialization v_0. However, unlike the low dimension case, here we cannot use Algorithm 2 to find such a initialization due to the assumption of $d \gg n$. Thus, we leave it as an open problem for privately finding such a initialization. Fortunately, in experiments we find randomly sample an initial vector can already achieve good performance.

Experimental studies: In Appendix, we provide empirical studies on the behaviors of our methods for (sparse) PCA, CCA and FDA on several real-world and synthetic data.

6 LOWER BOUNDS OF DP-GEP

In previous sections, we showed that for GEP in the CDP model under Assumption 1 and 2, it is possible to achieve an error of $\tilde{O}(\frac{d}{n} + \frac{d}{n^2\rho})$ and $\tilde{O}(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho})$ in low and high dimension sparse case under the statistical setting respectively. However, there are several questions left. First, can we further improve the error, i.e., what is the lower bound of error for GEP in the CDP model? Secondly, since all of our previous results are for the CDP or (ϵ, δ)-DP model. Thus, our question is, can we achieve similar results in the ϵ-DP model? In this section, we first show that the previous methods are near optimal for (sparse) PCA and (sparse) SIR in the CDP model. For the second one, we provide negative results by showing lower bounds of (sparse) PCA under the stochastic setting in ϵ-DP. Specifically, we show the following results.

Theorem 9 (Lower Bounds for Low Dimensional PCA)**.** For $0 < \epsilon \leq 1$, if $n \geq \Omega(\frac{d}{\epsilon})$, then for any ϵ-DP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{d}{n} + \frac{d^2}{n^2\epsilon^2}). \tag{23}$$

Moreover, for any $\rho > 1$, if $n \geq \Omega(\max\{d, \frac{\sqrt{d}}{\sqrt{\rho}}\})$, then for any ρ-CDP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), and

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{d}{n} + \frac{d}{n^2\rho}). \tag{24}$$

Here v^* is the leading eigenvector of $A = \mathbb{E}_{x\sim\mathcal{P}}[xx^T]]$.

Theorem 10 (Lower Bounds for High Dimensional Sparse PCA)**.** For $0 < \epsilon \leq 1$, if $n \geq \Omega(\frac{s\log d}{\epsilon})$, then for any ϵ-DP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$, if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), and its largest eigenvector v^* of the covariance matrix $A = \mathbb{E}_{x\sim\mathcal{P}}[xx^T]$ is s-sparse, such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{s\log d}{n} + \frac{(s\log d)^2}{n^2\epsilon^2}). \tag{25}$$

Moreover, for any $\rho > 0$, if n is sufficiently large such that $n \geq \Omega(\max\{s\log d, \frac{\sqrt{s\log d}}{\sqrt{\rho}}\})$, then for any ρ-DP algorithm with output v_{priv}, there exists a distribution $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$, if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), and its largest eigenvector v^* of the covariance matrix $A = \mathbb{E}_{x\sim\mathcal{P}}[xx^T]$ is s-sparse, and

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho}). \tag{26}$$

Next we consider the lower bounds of SIR in the CDP model, for simplicity we only consider the case where $k = 2$. That is we have two classes $Y = 1$ and $Y = 2$.

Theorem 11. For any $\rho > 0$, if $n \geq \Omega(\max\{d, \frac{\sqrt{d}}{\sqrt{\rho}}\})$, then for any ρ-CDP algorithm with output v_{priv}, there exists an instance $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{d}{n} + \frac{d}{n^2\rho}). \tag{27}$$

Here $\|v^*\|_2 = 1$ is the leading generalized eigenvector of the corresponding SIR.

Theorem 12. For any $\rho > 0$, if n is sufficiently large such that $n \geq \Omega(\max\{s\log d, \frac{\sqrt{s\log d}}{\sqrt{\rho}}\})$, then for any ρ-CDP

algorithm with output v_{priv}, there exists an instance $\mathcal{P}$ with $\mathbb{E}_{x\sim\mathcal{P}}[x] = 0$ and if $x_i \sim \mathcal{P}$ then it satisfies Assumption 1 and 2 (with high probability), such that

$$\mathbb{E}_{D\sim\mathcal{P}^n,\mathcal{A}}[1 - \frac{\langle v_{\text{priv}}, v^*\rangle}{\|v_{\text{priv}}\|_2}] \geq \Omega(\frac{s\log d}{n} + \frac{s\log d}{n^2\rho}). \quad (28)$$

Here $\|v^*\|_2 = 1$ is the leading generalized eigenvector of the corresponding sparse SIR with $\|v^*\|_0 \leq s$.

7 CONCLUSIONS

In this paper we provided the first study on the theoretical behaviors of the (sparse) Generalized Eigenvalue Problem (GEP) in the Differential Privacy (DP) model. Specifically, we considered both stochastic setting and deterministic setting in the low dimensional and high dimensional sparse cases. With some additional assumptions, we showed that our algorithms could achieve near optimal rates of error under the stochastic setting in both low dimensional and high dimensional sparse cases. Moreover, we provided the lower bound of (sparse) GEP in the ϵ-DP model to show a gap of the problem in the (ϵ, δ)-DP model.

However, there are still several unsolved problems left. First, from lower bounds and upper bounds of the error we can see that there is still a gap of $\text{Poly}(\log n)$ factor. Thus, can we further improve the upper bounds of error? Secondly, in the low dimension case, we discussed how to find an appropriate initial vector privately and efficiently. However, our approach cannot be extended to the high dimensional sparse case since we need to assume the sample size is large enough such that $n \gg d$, which violates the high dimension assumption. Thus, how do we find the initial vector privately in this case? Thirdly, for the lower bounds we proposed, we only considered the case for (sparse) PCA with sub-Gaussian distribution, where $\rho(E_A, k), \rho(E_B, k) = O(\sqrt{\frac{k\log d}{n}})$ and $\|E_A\|_2, \|E_B\|_2 = O(\sqrt{\frac{d}{n}})$. Thus, our question is, can we provide more general lower bounds which involve general $\rho(E_A, k)$ and $\rho(E_B, k)$? Finally, in the lower bound part we mainly focused on the stochastic setting. In the deterministic setting, [Dwork et al., 2014] provided the lower bound of PCA in the low dimension case. However, the lower bound of sparse PCA is still unknown. We will leave these open problems as future work.

ACKNOWLEDGEMENTS

Lijie Hu, Zihang Xiang and Di Wang are supported in part by the baseline funding BAS/1/1689-01-01, funding from the CRG grand URF/1/4663-01-01, FCC/1/1976-49-01 from CBRC and funding from the AI Initiative REI/1/4811-10-01 of King Abdullah University of Science and Technology (KAUST). Di Wang was also supported by the funding of the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI).

References

[Abadi et al., 2016] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

[Acharya et al., 2021] Acharya, J., Sun, Z., and Zhang, H. (2021). Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR.

[Balcan et al., 2016] Balcan, M.-F., Du, S. S., Wang, Y., and Yu, A. W. (2016). An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309.

[Barber and Duchi, 2014] Barber, R. F. and Duchi, J. C. (2014). Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*.

[Bassily et al., 2014] Bassily, R., Smith, A., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE.

[Biswas et al., 2020] Biswas, S., Dong, Y., Kamath, G., and Ullman, J. (2020). Coinpress: Practical private mean and covariance estimation. *arXiv preprint arXiv:2006.06618*.

[Blum et al., 2005] Blum, A., Dwork, C., McSherry, F., and Nissim, K. (2005). Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138. ACM.

[Bun and Steinke, 2016] Bun, M. and Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer.

[Cai et al., 2013] Cai, T. T., Ma, Z., Wu, Y., et al. (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110.

[Cai et al., 2019] Cai, T. T., Wang, Y., and Zhang, L. (2019). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*.

[Chaudhuri et al., 2012] Chaudhuri, K., Sarwate, A., and Sinha, K. (2012). Near-optimal differentially private principal components. *Advances in Neural Information Processing Systems*, 25:989–997.

[Chaudhuri et al., 2013] Chaudhuri, K., Sarwate, A. D., and Sinha, K. (2013). A near-optimal algorithm for differentially-private principal components. *The Journal of Machine Learning Research*, 14(1):2905–2943.

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

[Dwork et al., 2006] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

[Dwork et al., 2014] Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. (2014). Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20. ACM.

[Gao et al., 2015] Gao, C., Ma, Z., Ren, Z., and Zhou, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *The Annals of Statistics*, 43(5):2168–2197.

[Ge et al., 2018] Ge, J., Wang, Z., Wang, M., and Liu, H. (2018). Minimax-optimal privacy-preserving sparse pca in distributed systems. In *International Conference on Artificial Intelligence and Statistics*, pages 1589–1598.

[Golub and Van Loan, 1996] Golub, G. H. and Van Loan, C. F. (1996). Matrix computations. johns hopkins studies in the mathematical sciences.

[Gonem and Gilad-Bachrach, 2018] Gonem, A. and Gilad-Bachrach, R. (2018). Smooth sensitivity based approach for differentially private pca. In *Algorithmic Learning Theory*, pages 438–450.

[Hardt and Price, 2014] Hardt, M. and Price, E. (2014). The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869.

[Hardt and Roth, 2013] Hardt, M. and Roth, A. (2013). Beyond worst-case analysis in private singular vector computation. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 331–340. ACM.

[Hu et al., 2021] Hu, L., Ni, S., Xiao, H., and Wang, D. (2021). High dimensional differentially private stochastic optimization with heavy-tailed data. *arXiv preprint arXiv:2107.11136*.

[Imtiaz and Sarwate, 2017] Imtiaz, H. and Sarwate, A. D. (2017). Differentially-private canonical correlation analysis. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 283–287. IEEE.

[Imtiaz and Sarwate, 2019] Imtiaz, H. and Sarwate, A. D. (2019). Distributed differentially-private canonical correlation analysis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3112–3116. IEEE.

[Jin et al., 2019] Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. (2019). A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*.

[Kamath et al., 2021] Kamath, G., Liu, X., and Zhang, H. (2021). Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*.

[Kapralov and Talwar, 2013] Kapralov, M. and Talwar, K. (2013). On differentially private low rank approximation. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1395–1414. SIAM.

[Laurent and Massart, 2000] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

[Li, 1991] Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

[Liu and Altman, 2015] Liu, T. and Altman, R. B. (2015). Relating essential proteins to drug side-effects using canonical component analysis: a structure-based approach. *Journal of chemical information and modeling*, 55(7):1483–1494.

[Liu et al., 2022] Liu, X., Kong, W., Jain, P., and Oh, S. (2022). Dp-pca: Statistically optimal and differentially private pca. *arXiv preprint arXiv:2205.13709*.

[Massart, 2007] Massart, P. (2007). Concentration inequalities and model selection.

[Parkhomenko et al., 2009] Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical applications in genetics and molecular biology*, 8(1).

[Parlett, 1998] Parlett, B. N. (1998). *The symmetric eigenvalue problem*. SIAM.

[Shen, 2020] Shen, Y. (2020). Differentially private nonlinear canonical correlation analysis. In *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pages 1–5. IEEE.

[Stewart, 1979] Stewart, G. (1979). Pertubation bounds for the definite generalized eigenvalue problem. *Linear algebra and its applications*, 23:69–85.

[Strickert et al., 2009] Strickert, M., Keilwagen, J., Schleif, F.-M., Villmann, T., and Biehl, M. (2009). Matrix metric adaptation linear discriminant analysis of biomedical data. In *International Work-Conference on Artificial Neural Networks*, pages 933–940. Springer.

[Szarek, 1982] Szarek, S. J. (1982). Nets of grassmann manifold and orthogonal group. In *Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981)*, volume 169, page 185.

[Tan et al., 2018] Tan, K. M., Wang, Z., Liu, H., and Zhang, T. (2018). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1057–1086.

[Vershynin, 2009] Vershynin, R. (2009). On the role of sparsity in compressed sensing and random matrix theory. In *2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 189–192. IEEE.

[Vershynin, 2018] Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

[Vu et al., 2013a] Vu, V. Q., Cho, J., Lei, J., and Rohe, K. (2013a). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. *Advances in neural information processing systems*, 26.

[Vu et al., 2013b] Vu, V. Q., Lei, J., et al. (2013b). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947.

[Wang et al., 2019] Wang, D., Chen, C., and Xu, J. (2019). Differentially private empirical risk minimization with non-convex loss functions. In *International Conference on Machine Learning*, pages 6526–6535. PMLR.

[Wang and Xu, 2020] Wang, D. and Xu, J. (2020). Principal component analysis in the local differential privacy model. *Theoretical computer science*, 809:296–312.

[Wang et al., 2018] Wang, D., Ye, M., and Xu, J. (2018). Differentially private empirical risk minimization revisited: Faster and more general. *arXiv preprint arXiv:1802.05251*.

[Wang and Gu, 2020] Wang, L. and Gu, Q. (2020). A knowledge transfer framework for differentially private sparse learning. In *AAAI*, pages 6235–6242.

[Wu et al., 2017] Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. (2017). Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1307–1322.

[Yuan et al., 2019] Yuan, G., Shen, L., and Zheng, W.-S. (2019). A decomposition algorithm for the sparse generalized eigenvalue problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6113–6122.