Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Smooth Sensitivity Based Approach for Differentially Private Principal Component Analysis

Alon Gonen AGONEN@CS.PRINCETON.EDU
Department of Computer Science
Princeton University

Ran Gilad-Bachrach RANG@MICROSOFT.COM
Microsoft Research

Editor: Kevin Murphy and Bernhard Schölkopf

Abstract

We consider the challenge of differentially private PCA. Currently known methods for this task either employ the computationally intensive *exponential mechanism* or require an access to the covariance matrix, and therefore fail to utilize potential sparsity of the data. The problem of designing simpler and more efficient methods for this task has been raised as an open problem in Kapralov and Talwar (2013).

In this paper we address this problem by employing the output perturbation mechanism. Despite being arguably the simplest and most straightforward technique, it has been overlooked due to the large *global sensitivity* associated with publishing the leading eigenvector. We tackle this issue by adopting a *smooth sensitivity* based approach, which allows us to establish differential privacy (in a worst-case manner) and near-optimal sample complexity results under eigengap assumption. We consider both the pure and the approximate notions of differential privacy, and demonstrate a tradeoff between privacy level and sample complexity. We conclude by suggesting how our results can be extended to related problems.

1. Introduction

Differential Privacy has become a crucial requirement in many machine learning tasks involving private data such as medical and financial records (Dwork (2008); Dwork et al. (2014a); Chaudhuri et al. (2011); Blum et al. (2013); McSherry and Talwar (2007)). Informally speaking, a mechanism is said to be differentially private if one can hardly distinguish between two outputs of the algorithm corresponding to samples that differ in one entry. Since each entry typically corresponds to records of a single person, differential privacy essentially requires that the participation of a single individual in the sample (e.g. medical tests) would not reveal its private information. This requirement inherently implies a tradeoff between privacy and accuracy. Accordingly, considerable efforts have been made to identify structural properties that enable us to reduce this conflict.

Principal component analysis (PCA) is a fundamental dimensionality reduction technique in machine learning and data science. Finding a low-rank approximation of a given dataset is beneficial in terms of time and space complexity. In some scenarios (e.g. vision tasks), it also has the benefit of noise removal.

In view of the above, it is not surprising that differentially private PCA has received substantial attention recently (Blum et al. (2005); Chaudhuri et al. (2012); Kapralov and Talwar (2013); Hardt and Roth (2013); Dwork et al. (2014b)).

Our main contribution is a simple yet efficient method to make PCA differentially private. In a nut-shell, our method modifies standard PCA algorithm by adding a post-processing step in which a suitable noise is added to the output. Therefore, it is straightforward to combine it with any PCA implementation, including implementation that make use of unique properties of the data such as sparsity. To achieve that, we show that if there is a large eigengap between the leading eigenvalues of the covariance matrix, then the PCA problem becomes less sensitive to changes in its inputs. Hence, we can compute the amount of noise to inject as a function of the eigen gap.

1.1 Problem Definition

Let us now describe the considered problem formally. Let $\mathcal{D}$ an unknown distribution defined on the unit ball in $\mathbb{R}^d$.[1] Given a low-rank parameter $k \in [d]$, our ultimate goal is to approximately solve

$$\min_{U \in \mathcal{U}} F(U) = -\mathrm{tr}(UU^\top C), \text{ where } C := \mathbb{E}_{x \sim \mathcal{D}}[xx^\top], \;\; \mathcal{U} = \{U \in \mathbb{R}^{d \times k} : \; U^\top U = I_k\} \; , \quad (1)$$

while preserving differential privacy. The input to the learning algorithm $\mathcal{A}$ consists of a sample $S = (x_1, \ldots, x_n)$ drawn i.i.d. according to $\mathcal{D}$. Its output is denoted by $\hat{U} \in \mathcal{U}$. The sample complexity of the algorithm is a function $n : (0,1)^3 \to \mathbb{N}$, where $n(\epsilon_{\mathrm{g}}, \epsilon_{\mathrm{p}}, \delta_{\mathrm{p}})$ is the minimal size of an i.i.d. sample $S = (x_1, \ldots, x_n) \sim \mathcal{D}^n$ for which the following conditions simultaneously hold:

ϵ_{g}-accuracy: [2]with constant probability over both the draw of the sample S according to $\mathcal{D}^n$ and the internal randomness of the algorithm,

$$F(\hat{U}) \leqslant \min_{U \in \mathcal{U}} F(U) + \epsilon_{\mathrm{g}} \; .$$

$(\epsilon_{\mathrm{p}}, \delta_{\mathrm{p}})$-differential privacy: Let $d(S, S')$ be the minimal number of elements that should be removed or added to the sample S' to obtain the sample S. We say that S and S' are neighboring samples if $d(S, S') \leqslant 1$. We require that for all neighboring samples S, S', and for all $U \in \mathcal{U}$,

$$p(\mathcal{A}(S) = U) \leqslant \exp(\epsilon_{\mathrm{p}}) p(\mathcal{A}(S') = U) + \delta_{\mathrm{p}} \; , \quad (2)$$

(where p refers to the density function). The stricter notion of "pure" differential privacy requires also that $\delta_{\mathrm{p}} = 0$.

2. Algorithms and Main Result

In this paper we focus on particularly simple and efficiency-preserving method, named output perturbation. As its name suggests, the basic idea is to add noise to the output

[1] Our results can be easily scaled to balls of larger radius.

[2] Given a confidence parameter δ, standard techniques can be used to decrease the probability of failure to δ while incurring only logarithmic overhead in terms of sample complexity.

of an (approximately) exact algorithm. Arguably, this is the simplest and most flexible method, as it can be applied to any algorithm in a black-box fashion while preserving its efficiency. We assume an access to an algorithm $\mathcal{A}$ which (approximately) minimize the *empirical risk*

$$\hat{F}(U) := -\mathrm{tr}(UU^\top \hat{C}), \quad \hat{C} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \ .$$

We also assume that the algorithm $\mathcal{A}$ outputs the gap between the k-th and the $(k+1)$-th eigenvalues of $\hat{C}$. This assumption is not restrictive, as every reasonable PCA solver possesses this capability. Based on the output of $\mathcal{A}$, our mechanism determines the noise level. The main challenge in our work is to set the noise level so that differential privacy holds for any sample, and high accuracy is achieved under eigengap assumption.

Before adding the noise, there is another subtle issue which should be carefully addressed. To illustrate this challenge, consider the case $k = 1$. Clearly, a unit vector $u \in \mathbb{R}^d$ is a leading eigenvector if and only if $-u$ is also a leading eigenvector. Since the sign of the vector is arbitrary, a PCA solver might use it to leak private information, such as whether a specific point x^* was in the dataset ot not. Overcoming this potential risk is possible by negating the output of the PCA solver with probability 1/2 before adding the noise. More generally, for the case $k > 1$, we will replace $\hat{U}$ by $R\hat{U}$, where $R \in \mathbb{R}^{d \times d}$ is a random orthogonal matrix. We then add the noise and perform QR decomposition to obtain the final output. A detailed pseudocode of our method is given in Algorithm 1. To simplify the presentation and for the sake of conciseness, we focus on the case $k = 1$. The case of $k > 1$ is a straightforward extension of the case $k = 1$

Algorithm 1 Differentially private PCA using Output perturbation

```
Parameters: $\epsilon_{\mathrm{g}}, \delta_{\mathrm{g}}, \epsilon_{\mathrm{p}}, \delta_{\mathrm{p}} \in (0,1),\ k \in [d], \mathtt{PURE} \in \{\mathtt{TRUE}, \mathtt{FALSE}\}$
Input: $\hat{U} := \arg\min_{U \in \mathcal{U}} -\mathrm{tr}(U\hat{C}),\ \ G = \lambda_1(\hat{C}) - \lambda_2(\hat{C})$
Oracle: $\mathcal{A}(S) = (\tilde{U} := \arg\min_{U \in \mathcal{U}} \hat{F}(U),\ \lambda_k(\hat{C}) - \lambda_{k+1}(\hat{C}))$
Draw a random orthogonal matrix $R \in \mathbb{R}^{d \times d}$
Replace $\hat{U}$ with $\overline{U} = RU$
if $\mathtt{PURE} = \mathtt{TRUE}$ then
    Draw $E := E_{\mathtt{PURE}} \in \mathbb{R}^{d \times k}$ as described in Equation (4)
else
    Draw $E := E_{\mathtt{APPROX}} \in \mathbb{R}^{d \times k}$ as described in Equation (3)
end if
Return the matrix $\tilde{U} = QR(\overline{U} + E)$.
```

Theorem 1 ***(Main theorem: approximate case)*** *Given that* $\mathtt{PURE} = \mathtt{FALSE}$*, Algorithm 1 is* (ϵ_p, δ_p)*-differentially private. Furthermore, if* $\mathcal{GAP}(\mathcal{D}) := \lambda_1(\mathbb{E}[xx^\top]) - \lambda_2(\mathbb{E}[xx^\top]) > 0$*, then its sample complexity is at most*[3]

$$n(\epsilon_g, \epsilon_p, \delta_p) = \tilde{O}\left(\frac{\sqrt{d}}{\mathcal{GAP}(\mathcal{D})\epsilon_p\epsilon_g}\right)$$

[3] We use the $\tilde{O}$ notation to hide logarithmic dependencies.

Theorem 2 ***(Main theorem: pure case)*** *Given that* `PURE` $=$ `TRUE`*, Algorithm 1 is* ϵ_p*-differentially private. Furthermore, if* $\mathcal{GAP}(\mathcal{D}) := \lambda_1(\mathbb{E}[xx^\top]) - \lambda_2(\mathbb{E}[xx^\top]) > 0$*, then its sample complexity is at most*

$$n(\epsilon_g, \epsilon_p) = \tilde{O}\left(\frac{d^{3/2}}{\mathcal{GAP}(\mathcal{D})\epsilon_p\epsilon_g}\right)$$

3. Related Work

Differentially private PCA has been extensively investigated in Chaudhuri et al. (2012); Hardt and Roth (2013); Kapralov and Talwar (2013); Blum et al. (2005); Dwork et al. (2014b); Hardt and Roth (2012). The lower bound of Dwork et al. (2014b) implies that our sample complexity for the approximate case (see Theorem 1) is optimal up to logarithmic factors. For the pure case, the lower bound given by Chaudhuri et al. (2012) scales with d, whereas our upper bound (see Theorem 2) scales with $d^{3/2}$.

The first proposed method for differential private PCA was Sub-Linear Queries (SULQ) (Blum et al. (2005)). It employs the general strategy of *input perturbation* by adding random Gaussian noise to the empirical covariance matrix. Both the algorithm and its analysis have been refined recently by Dwork et al. (2014b). Restating their results within our framework gives approximate differential privacy with sample complexity bound identical to Theorem 1. They also consider the gap-free scenario. As we mentioned previously, the main limitation of this method is that it requires an access to the covariance matrix, which might be too costly in terms of space and time. Many fast PCA implementations (e.g. Shamir (2016); Ghashami et al. (2016); Clarkson and Woodruff (2012); Jain et al. (2016); Jin et al. (2015)) avoid working with the covariance matrix and consequently utilize potential sparsity of the data. As we mentioned previously, our output perturbation can be combined with any of these methods.

Another approach that has been investigated in Chaudhuri et al. (2012); Kapralov and Talwar (2013) is to use the exponential mechanism (Dwork et al. (2014a)). While this approach achieves pure differential with optimal sample complexity (also in the gap-free case), the only theoretically analyzed implementation of the associated sampling method runs in time $O(d^6)$.

Besides the spectral gap assumption, another common approach is to assume some form of incoherence. This route has been taken by Hardt and Roth (2013, 2012) who provide several interesting differentially private methods for PCA.

4. Analysis

In this section we prove our main result. We start by defining the local and global sensitivity of PCA, and proceed to define and analyze the smooth sensitivity.

4.1 Local and Global Sensitivity up To Equivalence

In the context of output perturbation, the sensitivity of a sample is defined as the maximum distance between two outputs of PCA corresponding to the neighboring samples. Unless specified otherwise, the distance is measured according to the ℓ_2-norm. Due to the equivalence between outputs discussed above, it makes sense to define the notion of distance between equivalent solutions. Namely, for any $U \in \mathcal{U}$, we define $[U] = \{RU : R \in \mathbb{R}^d, R^\top R = I_d\}$. The distance between $[U]$ and $[V]$ is defined by $\|[U] - [V]\| = \min\{\|U' - V'\| : U' \in [U], V' \in [V]\}$. Since our algorithm replaces the output $\mathcal{U}$ of PCA by $R\hat{U}$, where R is a random orthogonal matrix, this modification does not harm our analysis.

Definition 3 ***(Global and local sensitivity)*** *The ℓ_2-local sensitivity of a PCA algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{U}$ w.r.t. a sample $S = (x_1, \ldots, x_n)$ is defined as*

$$\mathcal{LS}(S) := \mathcal{LS}_{\mathcal{A}}(S) = \max_{S' : d(S,S') \leqslant 1} \|[\mathcal{A}(S)] - [\mathcal{A}(S')]\| \ .$$

The global sensitivity of $\mathcal{A}$ is defined as $\sup\{LS(S) : S \in \operatorname{supp}(\mathcal{D}^n)\}$. *The ℓ_1-local sensitivity is defined analogously.*

It is known that adding noise proportional to the global sensitivity (using a suitable noise distribution depending on the privacy parameters) yields differential privacy (Dwork et al. (2014b)). The following example due to Chaudhuri et al. (2012) illustrates the difficulty in preserving both accuracy and privacy using this approach. Let $u, u' \in \mathbb{R}^d$ be two orthonormal vectors and consider two samples S and S', where S consists of $n+1$ copies of u and n copies of u', whereas S' consists of $n+1$ copies of u' and n copies of u. The leading eigenvectors associated with S and S' are u and u', respectively. To satisfy differential privacy in this case, one should inject a noise proportional to the distance between u and u'. In particular, the amount of noise does not decreases as a function of the sample size, hence accuracy can not be preserved.

An easy computation shows that the eigengap in the previous examples scales inversely with the sample size. The following theorem shows that the larger the eigengap the smaller is the local sensitivity. We first make the following definition.

Definition 4 *Given a sample $S = (x_1, \ldots, x_n)$, we denote the eigengap between the two leading eigenvalues of the empirical covariance matrix $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ by $\mathcal{GAP}(S)$.*

Theorem 5 *Let $S = (x_1, \ldots, x_n) \in \operatorname{supp}(\mathcal{D}^n)$ be a sample and suppose $\mathcal{GAP}(S) > 0$. Then there exists a global constant $C > 0$ such that the ℓ_2-sensitivity of PCA is at most $\frac{3C}{n \cdot \mathcal{GAP}(S)}$. Furthermore, the global ℓ_2-sensitivity is $\sqrt{2}$. The ℓ_1-local sensitivity is at most $\sqrt{d}$ times larger than the ℓ_2-local sensitivity, and the ℓ_1-global sensitivity is at most 2.*

This result can be proved in several ways. The approach taken here exploits recent results on strict saddle problems, which include PCA as a special case.

Proof Let $S = (x_1, \ldots, x_n) \in \operatorname{supp}(\mathcal{D}^n)$ and $S' = (x_1, \ldots, x_{n-1}) \in \operatorname{supp}(\mathcal{D}^{n-1})$ be two neighboring samples and let u, u' be the minimizers of the corresponding empirical risks. Denote by $\hat{C} = n^{-1}\sum_{i=1}^n x_i x_i^\top$ and $\hat{C}' = n^{-1}\sum_{i=1}^n x_i x_i^\top$. By KKT conditions (Borwein et al. (2010)), there exist $\lambda := \lambda(u)$ and $\lambda' = \lambda(u')$ such that,

$$u = \arg\min_{v \in \mathbb{R}^d} \underbrace{-v^\top C v + \lambda(\|v\|^2 - 1)}_{=:\hat{L}(v)}, \quad u' = \arg\min_{v \in \mathbb{R}^d} \underbrace{-v^\top \hat{C}' v + \lambda'(\|v\|^2 - 1)}_{=:\hat{L}'(v)}$$

Also, λ and λ' admit the closed forms:

$$\lambda = u^\top \hat{C} u, \quad \lambda' = u'^\top \hat{C}' u' \ .$$

That is, λ is the leading eigenvalue of $\hat{C}$ and λ' is the leading eigenvalue of $\hat{C}'$. By first-order conditions, both $\nabla \hat{L}(u) = 0$ and $\nabla \hat{L}'(u') = 0$. Also,

$$\nabla \hat{L}(u') = -\hat{C}u' + \lambda u' = \frac{n-1}{n} \nabla \hat{L}'(u') - n^{-1} x_n x_n^\top u - \frac{n-1}{n} \lambda' u' + \lambda u' = n^{-1} x_n x_n^\top u - (\lambda' - \lambda) u + n^{-1} \lambda' u'$$

Since $\|x_i\| \leqslant 1$ for all i, by using Weyl's inequality we obtain that $\|\nabla \hat{L}(u')\| \leqslant \frac{3}{n}$.

We next use the strict saddle property of PCA to bound the distance between u and u'. Concretely, it is shown in Gonen and Shalev-Shwartz (2017) that our formulation of PCA is (α, γ, τ)-strict saddle with $\alpha, \gamma, \tau = C^{-1} \mathcal{GAP}(S)$ for some constant $C > 0$ (see Gonen and Shalev-Shwartz (2017)). Theorem 5 in this paper implies that if $n = \Omega(1/G^2)$, then u' lies in a $C^{-1} \mathcal{GAP}(S)$-strongly convex region (of the objective $\hat{F}$) around the minimizer u. By strong convexity (see Nesterov (2004)),

$$\nabla \hat{L}(u')^\top (u - u') = (\nabla \hat{L}(u') - \nabla \hat{L}(u))^\top (u - u') \geqslant C^{-1} \mathcal{GAP}(S) \|u - u'\|^2 \ .$$

Using Cauchy-Schwarz inequality, we obtain that

$$\|\nabla \hat{L}(u')\| \, \|u - u'\| \geqslant C^{-1} \mathcal{GAP}(S) \|u - u'\|^2 \Rightarrow \|u - u'\| \leqslant \frac{C}{\mathcal{GAP}(S)} \|\nabla \hat{L}(u')\| \leqslant \frac{3C}{n \mathcal{GAP}(S)} \ .$$

The bound on the ℓ_2-local sensitivity follows immediately. The bound on the ℓ_1-local sensitivity follows from the fact that the ℓ_1-distance is at most $\sqrt{d}$ larger than the ℓ_2-distance. The bounds on the global sensitivity are simply the ℓ_2 and the ℓ_1 distances between two perpendicular unit vectors. ■

It is tempting to replace the global sensitivity with the local one in hope of ensuring differential privacy in a worst case manner and achieving high accuracy under the common eigengap assumption. In general, this approach is problematic since the local sensitivity itself might be sensitive.[4] This brings us to the notion of *smooth sensitivity*, which we describe in the next part.

4.2 Background on Smooth sensitivity

Originated in (Nissim et al. (2007)), the notion of smooth sensitivity provides a systematic framework for generating insensitive surrogate for the local sensitivity. It consists of two main ingredients: a) Finding a suitable smooth upper bound on the local sensitivity. b) Generating noise according to an *admissible* distribution scaled by the smooth upper bound.

Definition 6 ***(smooth upper bound on the local sensitivity (Nissim et al. (2007)))*** *For $\beta > 0$, a function $\mathcal{SU} : \bigcup_{n \in \mathbb{N}} \mathrm{supp}(\mathcal{D}^n) \to \mathbb{R}_{\geqslant 0}$ is a β-smooth upper bound on the local sensitivity $\mathcal{LS}$ if it satisfies the following conditions:*

[4] The following example due to Dwork et al. (2014a) illustrates this idea. Suppose that we would like to compute the median of a given sequence in a differential private manner. Let S be a sample consisting of $n/2 + 1$ zeros and $n/2$ elements of magnitude at least 10^6. Assuming that we break ties in favor of the smaller value, the local sensitivity of S is zero. On the other hand, by removing a single zero element from S, we obtain a neighboring sample whose local sensitivity is at least 10^6.

1. *$\mathcal{SU}(S) \geqslant \mathcal{LS}(S)$ for every sample S*
2. *For every two neighboring samples S, S',*
$$\exp(-\beta)\mathcal{SU}(S') \leqslant \mathcal{SU}(S) \leqslant \exp(\beta)\mathcal{SU}(S') \ .$$

The following characterization of the smooth sensitivity is often useful.

Definition 7 *Let $\overline{\mathcal{LS}}$ be any point-wise upper bound on the local sensitivity. For a sample S and $k \in \mathbb{N}$, we define*
$$A^{(k)}(S) = \max_{S' \colon \ d(S,S') \leqslant k} \ \overline{\mathcal{LS}}(S') \ .$$

Lemma 8 ***(Nissim et al. (2007)[Claim 3.2])*** *Let $\overline{\mathcal{LS}}$ be any point-wise upper bound on the local sensitivity and define $A^{(k)}(S)$ as above. The function $U : \bigcup_{n \in \mathbb{N}} \operatorname{supp}(\mathcal{D}^n) \to \mathbb{R}$ defined by*
$$U(S) = \max_{k \in [n]} \exp(-\beta k) A^{(k)}(S)$$
is a β-smooth upper bound on the local sensitivity.

Analogously to the global sensitivity, the smooth local sensitivity determines the scale of the noise associated with our perturbed output. However, due to the change in the sensitivity level, the privacy guarantees are slightly worse than the standard case. For example, as explained in Nissim et al. (2007), drawing the noise according to any sub-gaussian distribution can not yield pure differential privacy. If one insists on obtaining pure differential privacy, more heavy-tailed distributions such as Cauchy distribution should be used. We discuss one non-pure (and less noisy) and one pure (and more noisy) possibilities.

Theorem 9 ***(Nissim et al. (2007)[Lemmas 2.7 and 2.10])***
Let $\epsilon_p, \delta_p \in (0,1)$ be the differential privacy parameters. The following claims hold:

1. ***Gaussian noise:*** *Suppose that $U(S)$ is a β-smooth upper bound on the local sensitivity, where $\beta = \frac{\epsilon_p}{4(d+\ln(2/\delta_p))}$. Define the noise matrix in Algorithm 1 by*
$$E_{\mathit{APPROX}} = \frac{5U(S) \cdot \sqrt{2\ln(2/\delta)}}{\epsilon_p} Z \ ,$$
where Z is a standard d-dimensional Gaussian random variable. Then Algorithm 1 is (ϵ_p, δ_p)-differentially private.

2. ***Cauchy noise:*** *Suppose that $U(S)$ is a β-smooth upper bound on the ℓ_1-local sensitivity, where $\beta = \frac{\epsilon_p}{6d}$. Define the noise matrix in Algorithm 1 by*
$$E_{\mathit{PURE}} = \frac{6U(S)}{\epsilon_p} Z \ ,$$
where $Z_1, \ldots, Z_d$ are drawn i.i.d. according to the density function $f(z) = \frac{1}{\pi(1+z^2)}$, is ϵ_p-differentially private.

Remark 10 *Lemmas 2.7 and 2.10 in Nissim et al. (2007) refer to the noisy output $\hat{U}$ before the QR step. However, since differential privacy is immune to post-processing (Dwork et al. (2014a)), the claim holds for the output $\tilde{U}$ as well.*

4.3 Smooth Sensitivity of PCA

In this part we bound the smooth sensitivity of PCA and establish the privacy properties of Algorithm 1.

Lemma 11 *Let $S = (x_1, \ldots, x_n)$ be a sample of size n and suppose that $\mathcal{GAP}(S) > 0$. For any sample $S' = (z_1, \ldots, z_m)$ with $d(S, S') \leqslant k$, we have that*

$$\max\{0, n \cdot \mathcal{GAP}(S) - k\} \leqslant m \cdot \mathcal{GAP}(S') \leqslant n \cdot \mathcal{GAP}(S) + k \ .$$

Furthermore, for each side of the above inequality, there exists a sample S' with $d(S, S') = k$ for which the inequality holds with equality.

Proof Let $H = \sum_{i=1}^n x_i x_i^\top$, $M = \sum_{i=1}^m z_i z_i^\top$ and denote by $P = H - M$. Using Weyl's inequality (Bhatia (1997)[Section 3.2]), we obtain that

$$\lambda_1(M) \geqslant \lambda_1(H) + \lambda_d(P), \quad \lambda_2(M) \leqslant \lambda_2(H) + \lambda_1(P)$$

Since the ℓ_2-norm of the x_i's (similarly, the z_i's) is at most 1 and $d(S, S') \leqslant k$, both the rank and the trace-norm of P are at most k.[5]Therefore,

$$\begin{aligned}\lambda_1(M) - \lambda_2(M) &\geqslant \lambda_1(H) - \lambda_2(H) + \lambda_d(P) - \lambda_1(P) \\ &\geqslant \lambda_1(H) - \lambda_2(H) - \sum_{i=1}^d |\lambda_i(P)| \\ &\geqslant \lambda_1(H) - \lambda_2(H) - k \ .\end{aligned}$$

This concludes the inequality. Letting u_2 be the second leading eigenvector of H, the right side of the inequality is attained by setting $S' = S + \sum_{i=1}^k u_2$. The left side is attained analogously. ■

Combining the last lemma with Theorem 5, Lemma 8 and Theorem 9, we conclude that Algorithm 1 is differentially private.

Corollary 12 ***(Approximate differential privacy)*** *Suppose that $\texttt{PURE} = \texttt{FALSE}$ and let*

$$E_{\texttt{PURE}} = \frac{5 \max_{k \in [n]} \exp(-\beta k) A^{(k)}(S) \cdot \sqrt{2 \ln(2/\delta)}}{\epsilon_p} Z \ , \tag{3}$$

where Z is standard d-dimensional Gaussian random variable and

$$A^{(k)}(S) = \begin{cases} \frac{C}{n \cdot \mathcal{GAP}(S) - k} & n \cdot \mathcal{GAP}(S) - k > 0 \\ \sqrt{2} & \textit{otherwise} \end{cases}$$

Then Algorithm 1 is (ϵ_p, δ_p)-differentially private.

[5] The trace norm of P is $\sum_{i=1}^d |\lambda_i(P)|$. Since P is the sum of k rank-1 matrices of trace 1, it follows using the triangle inequality that the trace norm of P is at most k.

Corollary 13 ***(Pure differential privacy)*** *Suppose that* `PURE` $=$ `TRUE` *and let*

$$E_{\texttt{PURE}} = \frac{6\max_{k\in[n]}\exp(-\beta k)A^{(k)}(S)}{\epsilon_p}Z\ , \tag{4}$$

where $Z_1,\dots,Z_d$ are i.i.d. Cauchy random variables and

$$A^{(k)}(S) = \begin{cases}\frac{C\sqrt{d}}{n\cdot\mathcal{GAP}(S)-k} & n\cdot\mathcal{GAP}(S)-k>0\\ 2 & \textit{otherwise}\end{cases}$$

Then Algorithm 1 is ϵ_p-differentially private.

4.4 Near-optimal accuracy under eigengap assumption

In this part we complete the proof of our main result by bounding the smooth sensitivity under the eigengap assumption. We start by relating the distributional gap assumption to the empirical eigengap. The following lemma follows from Matrix Bernstein inequality (Tropp and Others (2015)).

Lemma 14 *Suppose that $\mathcal{GAP}(\mathcal{D}) := \lambda_1(\mathbb{E}[xx^\top]) - \lambda_2(\mathbb{E}[xx^\top]) > 0$. If $n = \Omega\left(\frac{\log(d)}{\mathcal{GAP}(\mathcal{D})^2}\right)$, then with probability at least $1-\delta/2$ over the draw of a sample S according to $\mathcal{D}^n$, $\mathcal{GAP}(S) \geqslant \mathcal{GAP}(\mathcal{D})/2$.*

The next two lemma refer to the ℓ_1 and the ℓ_2 cases, respectively.

Lemma 15 *Let $\epsilon_p,\delta_p,\epsilon_g \in (0,1)$, and let $S=(x_1,\dots,x_n)\in \operatorname{supp}(\mathcal{D}^n)$ be a sample with $\mathcal{GAP}(S)>0$. Define $A^{(k)}(S)$ as in Corollary 12 and let $\beta = \frac{\epsilon_p}{4(d+\ln(2/\delta_p))}$. Suppose that*

$$n \geqslant \frac{2C\sqrt{d}}{\mathcal{GAP}(S)\epsilon_p\epsilon_g} + \frac{8(d+\ln(2/\delta_p))\ln(\sqrt{2}d/(\epsilon_p\epsilon_g))}{\mathcal{GAP}(S)\epsilon_p}\ .$$

Then

$$U(S) := \max_{k\in\mathbb{N}}\exp(-\beta k)A^{(k)}(S) \leqslant \epsilon_g\epsilon_p/\sqrt{d}$$

Proof Assume first that $k \leqslant \frac{n\cdot\mathcal{GAP}(S)}{2}$. In particular, this implies that $n\cdot\mathcal{GAP}(S)-k>0$. Using that $n \geqslant \frac{2C\sqrt{d}}{\mathcal{GAP}(S)\epsilon_{\mathrm{p}}\epsilon_{\mathrm{g}}}$, it follows that

$$\begin{aligned}U(S) := \exp(-\beta k)A^{(k)}(S) \leqslant A^{(k)}(S) &= \frac{C}{n\cdot\mathcal{GAP}(S)-k}\\ &\leqslant \frac{2C}{n\cdot\mathcal{GAP}(S)} \leqslant \epsilon_{\mathrm{g}}\epsilon_{\mathrm{p}}/\sqrt{d}\ .\end{aligned}$$

Assume now that $k > \frac{n\cdot\mathcal{GAP}(S)}{2}$, so $\exp(-\beta k) \leqslant \exp\left(-\frac{\beta n\cdot\mathcal{GAP}(S)}{2}\right)$. Using that $n \geqslant \frac{8(d+\ln(2/\delta_{\mathrm{p}}))\ln(\sqrt{2}d/(\epsilon_{\mathrm{p}}\epsilon_{\mathrm{g}}))}{\mathcal{GAP}(S)\epsilon_{\mathrm{p}}}$, we obtain that $\exp(-\beta k) \leqslant \epsilon_{\mathrm{g}}\epsilon_{\mathrm{p}}/(\sqrt{2}d)$ Since $A^{(k)} \leqslant \sqrt{2}$ for all S and k,

$$\exp(-\beta k)A^{(k)}(S) \leqslant \epsilon_{\mathrm{g}}\epsilon_{\mathrm{p}}/\sqrt{d}\ .$$

■

We proceed to the ℓ_1-case.

Lemma 16 *Let $\epsilon_p, \epsilon_g \in (0,1)$ and let $S = (x_1, \ldots, x_n) \in \mathrm{supp}(\mathcal{D}^n)$ be a sample with $\mathcal{GAP}(S) > 0$. Define $A^{(k)}(S)$ as in Corollary 13 and let $\beta = \frac{\epsilon_p}{6d}$. Suppose that*

$$n \geqslant \frac{2Cd^{3/2}}{\mathcal{GAP}(S)\epsilon_p\epsilon_g} + \frac{6d\,\ln(2d/(\epsilon_p\epsilon_g))}{\mathcal{GAP}(S)\epsilon_p} \ .$$

Then

$$U(S) := \max_{k \in \mathbb{N}} \exp(-\beta k) A^{(k)}(S) \leqslant \epsilon_g\epsilon_p/d \ .$$

We finally conclude our main result.

Proof (of Theorem 1 and Theorem 2) The differential privacy of the algorithm was established in Corollary 12 and Corollary 13. We next prove the bounds on the sample complexity. All the bounds given below hold with constant probability.

In view of Lemma 14, we may assume that $\mathcal{GAP}(S)$ is of order $\mathcal{GAP}(\mathcal{D})$. Sample complexity bounds for PCA (Gonen and Shalev-Shwartz (2017); Blanchard et al. (2007)) show that for $n = \Omega\left(\frac{1}{\mathcal{GAP}(\mathcal{D})\epsilon_{\mathrm{g}}}\right)$, the true risk of any unit vector is $\epsilon_{\mathrm{g}}/4$-close to its empirical risk. Therefore, adopting the notation used in Algorithm 1, $F(\hat{u}) \leqslant \min F(u) + \epsilon_{\mathrm{g}}/2$. It is left to show that

$$\hat{F}(\tilde{u}) - \hat{F}(\hat{u}) \leqslant \epsilon_{\mathrm{g}}/2 \ .$$

For the case $k = 1$, the QR decomposition step amounts to normalizing the noisy vector $\bar{u}$. Therefore, it suffices to show bound the ℓ_2 norm of the noise vector $\bar{u} - \hat{u}$ by ϵ_{g}. For approximate differential privacy, standard concentration bounds give a bound of order $\sqrt{d}$ on the ℓ_2 norm of a standard d-dimensional Gaussian vector. Using Lemma 15, we conclude the bound. For the pure setting, it is known that the median of the absolute value of a Cauchy random variable is 1. Since the Cauchy distribution is 1-stable, the sum of d i.i.d. Cauchy random variables is also a standard Cauchy random variable scaled by d. Consequently, the ℓ_1 and the ℓ_2 norms of the corresponding vector can be bounded by d (with constant probability). The desired bound follows from Lemma 16. ■

5. Discussion

In this work we studied the problem of adding privacy properties to the commonly used PCA algorithm. We showed that we can add privacy as a post processing step to any PCA solver while maintaining good accuracy. Moreover, the post processing step is efficient and preserves the utility of the PCA algorithm. This is a significant improvement over previous results that are either not computationally efficient or otherwise require changes to the implementations of PCA solvers.

We believe that some of the techniques used in our paper may be beneficial for other related problems. For example, our approach can be applied to any strict saddle problem for which we are able to compute the expression $A^{(k)}(S)$ which controls the smooth sensitivity. Furthermore, our technique for overcoming symmetry between equivalent solutions can be applied to most known strict saddle problems such as low rank problems whose minima are unique up to rotation (Ge et al. (2017)).

Acknowledgements

We would like to thank Kunal Talwar for sharing his ideas and knowledge.

References

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 1997. ISBN 0387948465. doi: 10.1007/978-1-4612-0653-8. URL `http://cnx.org/content/col10048/latest//{%}5Cnpapers2://publication/uuid/8E4EF096-A7F9-4960-AA3C-B2B8439EFCF0`.

Gilles Blanchard, Olivier Bousquet, and Laurent Zwald. Statistical properties of kernel principal component analysis. *Mach. Learn.*, 66(2-3):259–294, 2007.

Avrim Blum, Cynthia Dwork, Frank D. McSherry, and Kobbi Nissim. Practical privacy: the SuLQ framework. In *Proc. twenty-fourth ACM SIGMOD-SIGACT-SIGART Symp. Princ. database Syst.*, volume 2, pages 128–138. ACM, 2005. ISBN 1-59593-062-0. doi: http://doi.acm.org/10.1145/1065167.1065184. URL `http://doi.acm.org/10.1145/1065167.1065184`.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *J. ACM*, 60(2):12, 2013.

Jonathan M Borwein, Adrian S Lewis, Jonathan M. Borwein, and Adrian S.Lewis. *Convex analysis and nonlinear optimization: theory and examples*. Springer Science & Business Media, 2010. ISBN 1441921273 9781441921277. URL `http://www.amazon.com/Convex-Analysis-Nonlinear-Optimization-Mathematics/dp/1441921273/ref=tmm{_}pap{_}title{_}0?ie=UTF8{&}qid=1384184182{&}sr=1-2`.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12(Mar):1069–1109, 2011.

Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *Adv. Neural Inf. Process. Syst.*, pages 989–997, 2012.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proc. forty-fifth Annu. ACM Symp. Theory Comput.*, pages 81–90. ACM, 2012. ISBN 9781450320290. doi: 10.1145/2488608.2488620. URL `http://arxiv.org/abs/1207.6365`.

Cynthia Dwork. Differential privacy: A survey of results. In *Int. Conf. Theory Appl. Model. Comput.*, pages 1–19. Springer, 2008.

Cynthia Dwork, Aaron Roth, and Others. The algorithmic foundations of differential privacy. *Found. Trends®in Theor. Comput. Sci.*, 9(3–4):211–407, 2014a.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proc. 46th Annu. ACM Symp. Theory Comput.*, pages 11–20. ACM, 2014b.

Rong Ge, Chi Jin, and Yi Zheng. No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis. *arXiv Prepr. arXiv1704.00708*, 2017.

Mina Ghashami, Edo Liberty, and Jeff M. Phillips. Efficient Frequent Directions Algorithm for Sparse Matrices. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*, pages 845–854, 2016. ISSN 9781450321389. doi: 10.1145/2939672.2939800. URL `http://dl.acm.org/citation.cfm?doid=2939672.2939800`.

Alon Gonen and Shai Shalev-Shwartz. Fast Rates for Empirical Risk Minimization of Strict Saddle Problems. In Ohad Shamir and Satyen Kale, editors, *Proc. 2017 Conf. Learn. Theory*, pages 1043—–1063. PMLR, 2017. URL `http://arxiv.org/abs/1701.04271`.

Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *Proc. forty-fourth Annu. ACM Symp. Theory Comput.*, pages 1255–1268. ACM, 2012. ISBN 9781450312455. doi: 10.1145/2213977.2214088. URL `http://dl.acm.org/citation.cfm?doid=2213977.2214088`.

Moritz Hardt and Aaron Roth. Beyond worst-case analysis in private singular vector computation. In *Proc. forty-fifth Annu. ACM Symp. Theory Comput.*, pages 331–340. ACM, 2013. ISBN 9781450320290. doi: 10.1145/2488608.2488650. URL `http://dl.acm.org/citation.cfm?doid=2488608.2488650`.

Prateek Jain, Chi Jin, Sham M.Kakade, Praneeth Netrapalli, and Aaron Sidford. Streaming PCA : Matching Matrix Bernstein and Near-Optimal Finite Sample Guarantees for Oja ' s Algorithm. *JMLRWorkshop Conf. Proc.*, 49:1–18, 2016.

Chi Jin, Sham M. Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Robust Shift-and-Invert Preconditioning: Faster and More Sample Efficient Algorithms for Eigenvector Computation. *arXiv Prepr. arXiv1510.08896*, pages 1–38, 2015. URL `http://arxiv.org/abs/1510.08896`.

Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In *Proc. Twenty-Fourth Annu. ACM-SIAM Symp. Discret. Algorithms*, pages 1395–1414. SIAM, 2013. ISBN 9781611972511. doi: 10.1137/1.9781611973105.101. URL `http://dblp.uni-trier.de/db/conf/soda/soda2013.html{#}KapralovT13`.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Found. Comput. Sci. 2007. FOCS'07. 48th Annu. IEEE Symp.*, pages 94–103. IEEE, 2007.

Yurii Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004. ISBN 9781461346913.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. *Proc. thirty-ninth Annu. ACM Symp. Theory Comput. - STOC '07*, 1(x):75, 2007. ISSN 07378017. doi: 10.1145/1250790.1250803. URL `http://portal.acm.org/citation.cfm?doid=1250790.1250803`.

Ohad Shamir. Convergence of Stochastic Gradient Descent for PCA. *Int. Conf. Mach. Learn.*, 2016. URL `http://arxiv.org/abs/1509.09002`.

Joel A Tropp and Others. An introduction to matrix concentration inequalities. *Found. Trends{®} Mach. Learn.*, 8(1-2):1–230, 2015. ISSN 1935-8237. doi: 10.1561/2200000048.