DP-PCA: Statistically Optimal and Differentially Private PCA

Xiyang Liu
Paul Allen School of Computer Science & Engineering
University of Washington
xiyangl@cs.washington.edu

Weihao Kong
Google Research
weihaokong@google.com

Prateek Jain
Google Research
prajain@google.com

Sewoong Oh
Paul Allen School of Computer Science & Engineering
University of Washington, and Google Research
sewoong@cs.washington.edu

Abstract

We study the canonical statistical task of computing the principal component from n i.i.d. data in d dimensions under (ε, δ)-differential privacy. Although extensively studied in literature, existing solutions fall short on two key aspects: (i) even for Gaussian data, existing private algorithms require the number of samples n to scale super-linearly with d, i.e., $n = \Omega(d^{3/2})$, to obtain non-trivial results while non-private PCA requires only $n = O(d)$, and (ii) existing techniques suffer from a non-vanishing error even when the randomness in each data point is arbitrarily small. We propose DP-PCA, which is a single-pass algorithm that overcomes both limitations. It is based on a private minibatch gradient ascent method that relies on *private mean estimation* to add minimal noise required to ensure privacy by adapting to the geometry of a given minibatch of gradients. For sub-Gaussian data, we provide nearly optimal statistical error rates even for $n = \tilde{O}(d)$. Furthermore, we provide a lower bound showing that sub-Gaussian style assumption is necessary in obtaining the optimal error rate.

1 Introduction

Principal Component Analysis (PCA) is a fundamental statistical tool with multiple applications including dimensionality reduction, data visualization, and noise reduction. Naturally, it is a key part of most standard data analysis and ML pipelines. However, when applied to data collected from numerous individuals, such as the U.S. Census data, outcome of PCA might reveal highly sensitive personal information. We investigate the design of privacy preserving PCA algorithms and the involved privacy/utility tradeoffs, for computing the first principal component, that should serve as the building block of more general rank-k PCA.

Differential privacy (DP) is a widely accepted mathematical notion of privacy introduced in [21], which is a standard in releasing the U.S. Census data [2] and also deployed in commercial systems [72, 25, 27]. A query to a database is said to be (ε, δ)-differentialy private if a strong adversary who knows all other entries but one cannot infer that one entry from the query output, with high confidence. The parameters ε and δ restricts the confidence as measured by the Type-I and II errors [49]. Smaller values of $\varepsilon \in [0, \infty)$ and $\delta \in [0, 1]$ imply stronger privacy and plausible deniability for the participants.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

For non-private PCA with n i.i.d. samples in d dimensions, the popular Oja's algorithm (provided in Algorithm 1) achieves the optimal error of $\sin(\hat{v}, v_1) = \tilde{\Theta}(\sqrt{d/n})$, where the error is measured by the sine function of the angle between the estimate, $\hat{v}$, and the principal component, v_1, [45]. For differentially private PCA, there is a natural fundamental question: *what is the extra cost we pay in the error rate for ensuring* (ε, δ)*-DP?*

We introduce a novel approach we call DP-PCA (Algorithm 3) and show that it achieves an error bounded by $\sin(\hat{v}, v) = \tilde{O}(\sqrt{d/n} + d/(\varepsilon n))$ for *sub-Gaussian-like* data defined in Assumption 1, which is a broad class of light-tailed distributions that includes Gaussian data as a special case. The second term characterizes the cost of privacy and this is tight; we prove a nearly matching information theoretic lower bound showing that no (ε, δ)-DP algorithm can achieve a smaller error. This significantly improves upon a long line of existing private algorithms for PCA, e.g., [14, 9, 38, 36, 23]. These existing algorithms are analyzed for fixed and non-stochastic data and achieve sub-optimal error rates of $O(\sqrt{d/n} + d^{3/2}/(\varepsilon n))$ even in the stochastic setting we consider.

A remaining question is whether the sub-Gaussian-like assumption, namely Assumption A.4, is necessary or if it is an artifact of our analysis and our algorithm. It turns out that such an assumption on the lightness of the tail is critical; we prove an algorithmic independent and information theoretic lower bound (Theorem 5.4) to show that, without such an assumption, the cost of privacy is lower bounded by $\Omega(\sqrt{d/(\varepsilon n)})$. This proves a separation of the error depending on the lightness of the tail.

We start with the formal description of the stochastic setting in Section 2 and present Oja's algorithm for non-private PCA. Our first attempt in making this algorithm private in Section 3 already achieves near-optimal error, if the data is strictly from a Gaussian distribution. However, there are two remaining challenges that we describe in detail in Section 4: (i) the excessive number of iterations of Private Oja's Algorithm (Algorithm 2) prevents using typical values of ε used in practice, and (ii) for general sub-Gaussian-like distributions, the error does not vanish even when the noise in the data (as measured by a certain fourth moment of a function of the data) vanishes. The first challenge is due to the analysis that requires amplification by shuffling [24] that is restrictive. The second is due to its reliance on gradient norm clipping [1] that does not adapt to the geometry of the current gradients. This drives the design of DP-PCA in Section 5 that critically relies on two techniques to overcome each challenge, respectively. First, minibatch SGD (instead of single sample SGD) significantly reduces the number iterations, thus obviating the need for amplification by shuffling. Next, private mean estimation (instead of gradient norm clipping and noise adding) adapts to the geometry of the problem and adds the minimal noise necessary to achieve privacy. The main idea of this geometry adaptive stochastic gradient update is explained in detail in Section 6, along with a sketch of a proof.

Notations. For a vector $x \in \mathbb{R}^d$, we use $\|x\|$ to denote the Euclidean norm. For a matrix $X \in \mathbb{R}^{d\times d}$, we use $\|X\|_2 = \max_{\|v\|=1} \|Xv\|_2$ to denote the spectral norm. We use $\mathbf{I}_d$ to denote $d \times d$ identity matrix. For $n \in \mathbb{Z}^+$, let $[n] := \{1, 2, \ldots, n\}$. Let $\mathbb{S}_2^{d-1}$ denote the unit d-sphere of ℓ_2, i.e., $\mathbb{S}_2^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$. $\tilde{O}()$ hides logarithmic factors in n, d, and the failure probability ζ.

2 Problem formulation and background on DP

Typical PCA assumes i.i.d. data $\{x_i \in \mathbb{R}^d\}$ from a distribution and finds the first eigenvector of $\Sigma = \mathbb{E}[(x_i - \mathbb{E}[x_i])(x_i - \mathbb{E}[x_i])^\top] \in \mathbb{R}^{d\times d}$. Our approach allows for a more general class of data $\{A_i \in \mathbb{R}^{d\times d}\}$ that recovers the standard case when $A_i = (x_i - \mathbb{E}[x_i])(x_i - \mathbb{E}[x_i])^\top$.

Assumption 1 ($(\Sigma, \{\lambda_i\}_{i=1}^d, M, V, K, \kappa, a, \gamma^2)$-model)**.** *Let $A_1, A_2, \ldots, A_n \in \mathbb{R}^{d\times d}$ be a sequence of (not necessarily symmetric) matrices sampled independently from the same distribution that satisfy the following with PSD matrices $\Sigma \in \mathbb{R}^{d\times d}$ and $H_u \in \mathbb{R}^{d\times d}$, and positive scalar parameters M, V, K, κ, a, and γ^2:*

A.1. *Let $\Sigma := \mathbb{E}[A_i]$, for a symmetric positive semidefinite (PSD) matrix $\Sigma \in \mathbb{R}^{d\times d}$, λ_i denote the i-th largest eigenvalue of Σ, and $\kappa := \lambda_1/(\lambda_1 - \lambda_2)$,*

A.2. $\|A_i - \Sigma\|_2 \le \lambda_1 M$ *almost surely,*

A.3. $\max\left\{\left\|\mathbb{E}\left[(A_i - \Sigma)(A_i - \Sigma)^\top\right]\right\|_2, \left\|\mathbb{E}\left[(A_i - \Sigma)^\top(A_i - \Sigma)\right]\right\|_2\right\} \le \lambda_1^2 V,$

A.4. $\max_{\|u\|=1,\|v\|=1} \mathbb{E}\left[\exp\left(\left(\frac{|u^\top(A_i^\top-\Sigma)v|^2}{K^2\lambda_1^2\|H_u\|_2}\right)^{1/(2a)}\right)\right] \leq 2$, *where* $H_u := (1/\lambda_1^2)\mathbb{E}[(A_i - \Sigma)uu^\top(A_i-\Sigma)^\top]$. *We denote* $\gamma^2 := \max_{\|u\|=1}\|H_u\|_2$.

The first three assumptions are required for PCA even if privacy is not needed. The last assumption provides a Gaussian-like tail bound that determines how much noise we need to introduce in the algorithm for (ε,δ)-DP. The following lemma is useful in the analyses.

Lemma 2.1. *Under A.1 and A.4 in Assumption 1, for any unit vector u, v, with probability* $1-\zeta$,

$$|u^\top(A_i^\top-\Sigma)v|^2 \;\leq\; K^2\lambda_1^2\|H_u\|_2\log^{2a}(2/\zeta)\;. \tag{1}$$

2.1 Oja's algorithm

In a non-private setting, the following streaming algorithm introduced in [69] achieves optimal sample complexity as analyzed in [45]. It is a projected stochastic gradient ascent on the objective defined on the empirical covariance: $\max_{\|w\|=1}(1/n)\sum_{i=1}^n w^\top A_i w$.

Algorithm 1: (Non-private) Oja's Algorithm

Choose w_0 uniformly at random from the unit sphere
for $t=1,2,\ldots,T$ **do** $w_t' \leftarrow w_{t-1}+\eta_t A_t w_{t-1}$, $w_t \leftarrow w_t'/\|w_t'\|$
Return w_T

Central to our analysis is the following error bound on Oja's Algorithm from [45].

Theorem 2.2 ([45, Theorem 4.1])**.** *Under Assumptions A.1-A.3, suppose the step size* $\eta_t = \frac{\alpha}{(\lambda_1-\lambda_2)(\xi+t)}$ *for some* $\alpha > 1/2$ *and* $\xi := 20\max\left(\kappa M\alpha, \kappa^2(V+1)\alpha^2/\log(1+(\zeta/100))\right)$. *If* $T>\xi$ *then there exists a constant* $C>0$ *such that Algorithm 1 outputs* w_T *achieving w.p.* $1-\zeta$,

$$\sin^2(w_T,v_1) \leq \frac{C\log(1/\zeta)}{\zeta^2}\left(\frac{\alpha^2\kappa^2 V}{(2\alpha-1)T}+d\left(\frac{\xi}{T}\right)^{2\alpha}\right)\;. \tag{2}$$

2.2 Background on Differential Privacy

Differential privacy (DP), introduced in [21], is a de facto mathematical measure for privacy leakage of a database accessed via queries. It ensures that even an adversary who knows all other entries cannot identify with a high confidence whether a person of interest participated in a database or not.

Definition 2.3 (Differential privacy [21])**.** *Given two multisets* S *and* S', *we say the pair* (S,S') *is* neighboring *if* $|S\setminus S'|+|S'\setminus S|\leq 1$. *We say a stochastic query* q *over a dataset* S *satisfies* (ε,δ)-differential privacy *for some* $\varepsilon>0$ *and* $\delta\in(0,1)$ *if* $\mathbb{P}(q(S)\in A)\leq e^\varepsilon\mathbb{P}(q(S')\in A)+\delta$ *for all neighboring* (S,S') *and all subset* A *of the range of* q.

Small values of ε and δ ensures that the adversary cannot identify any single data point with high confidence, thus providing plausible deniability. We provide useful DP lemmas in Appendix B. Within our stochastic gradient descent approach to PCA, we rely on the Gaussian mechanism to privatize each update. The *sensitivity* of a query q is defined as $\Delta_q := \sup_{\text{neighboring }(S,S')}\|q(S)-q(S')\|$.

Lemma 2.4 (Gaussian mechanism [22])**.** *For a query* q *with sensitivity* Δ_q, $\varepsilon\in(0,1)$, *and* $\delta\in(0,1)$, *the Gaussian mechanism outputs* $q(S)+\mathcal{N}(0,(\Delta_q(\sqrt{2\log(1.25/\delta)})/\varepsilon)^2\mathbf{I}_d)$ *and achieves* (ε,δ)*-DP.*

This is a special case of a family of output perturbation mechanisms which includes the Laplace mechanism [21] and stair-case mechanisms [34]. The latter is shown to be optimal in one-dimension [35] and for hypothesis testing under local DP [48]. Another mechanism we frequently use is the private histogram learner of [56], whose analysis is provide in Appendix B, along with various composition theorems to provide end-to-end guarantees.

2.3 Comparisons with existing results in private PCA

We briefly discuss the most closely related work and provide more previous work in Appendix A. Most existing results assume a fixed data under a deterministic setting where each sample has a bounded

Algorithm 2: Private Oja's Algorithm

Input: $S = \{A_i \in \mathbb{R}^{d\times d}\}_{i=1}^n$, privacy (ε, δ), learning rates $\{\eta_t\}_{t=1}^n$

1 Randomly permute S and choose w_0 uniformly at random from the unit sphere

2 Set DP noise multiplier: $\alpha \leftarrow C' \log(n/\delta)/(\varepsilon\sqrt{n})$

3 Set clipping threshold: $\beta \leftarrow C\lambda_1\sqrt{d}(K\gamma \log^a(nd/\zeta) + 1)$

4 **for** *t=1, 2, . . . , n* **do**

5   Sample $z_t \sim \mathcal{N}(0, \mathbf{I}_d)$

6     $w'_t \leftarrow w_{t-1} + \eta_t \operatorname{clip}_\beta(A_t w_{t-1}) + 2\eta_t\beta\alpha z_t$ where $\operatorname{clip}_\beta(x) = x \cdot \min\{1, \frac{\beta}{\|x\|_2}\}$

7     $w_t \leftarrow w'_t/\|w'_t\|$

8 Return w_n

norm, $\|x_i\| \le \beta$, and the goal is to find the top eigenvector of $\hat\Sigma := (1/n)\sum_{i=1}^n (x_i - \hat\mu)(x_i - \hat\mu)^\top$ for the empirical mean $\hat\mu$. For the purpose of comparisons, consider Gaussian $x_i \sim \mathcal{N}(0, \Sigma)$ with $\|x_i\| \le \beta = O(\sqrt{\lambda_1 d \log(n/\zeta)})$ for all $i \in [n]$ with probability $1 - \zeta$. The first line of approaches in [9, 14, 23] is a Gaussian mechanism that outputs $\mathrm{PCA}(\widehat\Sigma + Z)$, where Z is a symmetric matrix with i.i.d. Gaussian entries with a variance $((\beta^2/n\varepsilon)\sqrt{2\log(1.25/\delta)})^2$ to ensure (ε, δ)-DP. The tightest result in [23, Theorem 7] achieves

$$\sin(\hat v, v_1) \quad = \quad \tilde O\Big(\kappa\Big(\sqrt{\frac{d}{n}} + \frac{d^{3/2}\sqrt{\log(1/\delta)}}{\varepsilon n}\Big)\Big)\,, \tag{3}$$

with high probability, under a strong assumption that the spectral gap is very large: $\lambda_1 - \lambda_2 = \omega(d^{3/2}\sqrt{\log(1/\delta)}/(\varepsilon n))$. In a typical scenario with $\lambda_1 = O(1)$, this requires a large sample size of $n = \omega(d^{3/2}/\varepsilon)$. Since this Gaussian mechanism does not exploit the statistical properties of i.i.d. samples, the second term in this upper bound is larger by a factor of $d^{1/2}$ compared to the proposed DP-PCA (Corollary 5.2). The error rate of Eq. (3) is also achieved in [38, 36] by adding Gaussian noise to the standard power method for computing the principal components. When the spectral gap, $\lambda_1 - \lambda_2$, is smaller, it is possible to trade-off the dependence in κ and the sampling ratio d/n, which we do not address in this work but is surveyed in Appendix A.

3 First attempt: making Oja's Algorithm private

Following the standard recipe in training with DP-SGD, e.g., [1] and a recent work [75], we introduce Private Oja's Algorithm in Algorithm 2. At each gradient update, we first apply gradient norm clipping to limit the contribution of a single data point and next add an appropriately chosen Gaussian noise from Lemma 2.4 to achieve (ε, δ)-DP, end-to-end. The choice of clipping threshold β ensures that, with high probability under our assumption, we do not clip any gradients. The choice of noise multiplier α ensures (ε, δ)-DP.

One caveat in streaming algorithms is that we access data n times, each with a private mechanism, but accessing only a single data point at a time. To prevent excessive privacy loss due to such a large number of data accesses, we apply a random shuffling in line 1 Algorithm 2, in order to benefit from a standard amplification by shuffling [24, 29]. This gives an amplified privacy guarantee that allows us to add a small noise proportional to $\alpha = O(\log(n/\delta)/(\varepsilon\sqrt{n}))$. Without the shuffle amplification, we will instead need a larger noise scaling as $\alpha = O(\log(n/\delta)/\varepsilon)$, resulting in a suboptimal utility guarantee. However, this comes with a restriction that the amplification holds only for small values of $\varepsilon = O(\sqrt{\log(n/\delta)/n})$. Our first contribution in the proposed DP-PCA (Algorithm 3) is to expand this range to $\varepsilon = O(1)$, which includes the practical regime of interest $\varepsilon \in [1/2, 5]$.

Lemma 3.1 (Privacy)**.** *If $\varepsilon = O(\sqrt{\log(n/\delta)/n})$ and the noise multiplier is chosen to be $\alpha = \Omega\left(\log(n/\delta)/(\varepsilon\sqrt{n})\right)$, then Algorithm 2 is (ε, δ)-DP.*

Under Assumption 1, we select gradient norm clipping threshold β such that no gradient exceeds β.

Lemma 3.2 (Gradient clipping)**.** *Let $\beta = C\lambda_1\sqrt{d}(K\gamma\log^a(nd/\zeta) + 1)$ for some constant $C > 0$. Then with probability $1 - \zeta$, $\|A_t w_{t-1}\| \le \beta$ for any fixed w_{t-1} independent of A_t, for all $t \in [n]$.*

We provide proofs of both lemmas and the next theorem in Appendix D. When no clipping is applied, we can use the standard analysis of Oja's Algorithm from [45] to prove the following utility guarantee.

Theorem 3.3 (Utility)**.** *Given n i.i.d. samples $\{A_i \in \mathbb{R}^{d\times d}\}_{i=1}^n$ satisfying Assumption 1 with parameters $(\Sigma, M, V, K, \kappa, a, \gamma^2)$, if*

$$n \;=\; \tilde{O}\Big(\kappa^2 + \kappa M + \kappa^2 V + \frac{d\,\kappa\,(\gamma+1)\,\log(1/\delta)}{\varepsilon}\Big)\,, \tag{4}$$

with a large enough constant, then there exists a positive universal constant c_1 and a choice of learning rate η_t that depends on $(t, M, V, K, a, \lambda_1, \lambda_1 - \lambda_2, n, d, \varepsilon, \delta)$ such that Algorithm 2 with a choice of $\zeta = 0.01$ outputs w_n that achieves with probability 0.99,

$$\sin^2(w_n, v_1) \;=\; \widetilde{O}\left(\kappa^2\Big(\frac{V}{n} + \frac{(\gamma+1)^2 d^2 \log^2(1/\delta)}{\varepsilon^2 n^2}\Big)\right)\,, \tag{5}$$

where $\widetilde{O}(\cdot)$ hides poly-logarithmic factors in n, d, $1/\varepsilon$, and $\log(1/\delta)$ and polynomial factors in K.

The first term in Eq. (5) matches the non-private error rate for Oja's algorithm in Eq. (2) with $\alpha = O(\log n)$ and $T = n$, and the second term is the price we pay for ensuring (ε, δ)-DP.

Remark 3.4. *For a canonical setting of a Gaussian data with $A_i = x_i x_i^\top$ and $x_i \sim \mathcal{N}(0, \Sigma)$, we have, for example from [70, Lemma 1.12], that $M = O(d\log(n))$, $V = O(d)$, $K = 4$, $a = 1$, and $\gamma^2 = O(1)$. Theorem 3.3 implies the following error rate:*

$$\sin^2(w_n, v_1) \;=\; \tilde{O}\Big(\kappa^2\Big(\frac{d}{n} + \frac{d^2\log^2(1/\delta)}{\varepsilon^2 n^2}\Big)\Big)\,. \tag{6}$$

Comparing to a lower bound in Theorem 5.3, this is already near optimal. However, for general distributions satisfying Assumption 1, Algorithm 2 (in particular the second term in Eq. (5)) can be significantly sub-optimal. We explain this second weakness of Private Oja's Algorithm in the following section (the first weakness is the restriction on $\varepsilon = O(\sqrt{\log(n/\delta)/n})$).

4 Two remaining challenges

We explain the two remaining challenges in Private Oja's Algorithm and propose techniques to overcomes these challenges that lead to the design of DP-PCA (Algorithm 3).

First challenge: restricted range of $\varepsilon = O(\sqrt{\log(n/\delta)/n})$. This is due to the large number, n, of iterations that necessitates the use of the amplification by shuffling in Theorem D.1. We reduce the number of iterations with minibatch SGD. For $T = O(\log^2 n)$ and $t = 1, 2, \ldots, T$, we repeat

$$w'_t \leftarrow w_{t-1} + \frac{\eta_t}{B}\sum_{i=1+B(t-1)}^{Bt-1} \mathrm{clip}_\beta(A_i w_{t-1}) + \frac{w\eta_t\beta\alpha}{B} z_t\ , \text{ and } \quad w_t \leftarrow w'_t/\|w'_t\|\ , \tag{7}$$

where $z_t \sim \mathcal{N}(0, \mathbf{I}_d)$ and the minibatch size is $B = \lfloor n/T \rfloor$. Since the dataset is accessed only $T = O(\log^2 n)$ times, the end-to-end privacy is analyzed with the serial composition (Lemma B.3) instead of the amplification by shuffling. This ensures (ε, δ)-DP for any $\varepsilon = O(1)$, resolving the first challenge, and still achieves the utility guarantee of Eq. (5).

Second challenge: excessive noise for privacy. This is best explained with an example.

Example 4.1 (Signal and noise separation)**.** *Consider a setting with $A_i = x_i x_i^\top$ and $x_i = s_i + n_i$ where $s_i = v$ with probability half and $s_i = -v$ otherwise for a unit norm vector v and $n_i \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. We want to find the principal component of $\Sigma = \mathbb{E}[x_i x_i^\top] = vv^\top + \sigma^2\mathbf{I}$, which is v. This construction decomposes the signal and the noise. For $A_i = vv^\top + s_i n_i^\top + n_i s_i^\top + n_i n_i^\top$, the signal component is determined by $vv^\top$ that is deterministic due to the sign cancelling. The noise component is $s_i n_i^\top + n_i s_i^\top + n_i n_i^\top$ which is random. We can control the Signal-to-Noise Ratio (SNR), $1/\sigma^2$, by changing σ^2, and we are particularly interested in the regime where σ^2 is small. As we are interested in $\sigma^2 < 1$, this satisfies Assumption 1 with $\lambda_1 = 1 + \sigma^2$, $\lambda_2 = \sigma^2$, $V = O(d\sigma^2)$, $K = O(1)$, $a = 1$, and $\gamma^2 = \sigma^2$. Substituting this into Eq.* (5)*, Private Oja's Algorithm achieves*

$$\sin^2(w_n, v_1) \;=\; \tilde{O}\Big(\frac{\sigma^2 d}{n} + \frac{d^2\log(1/\delta)}{\varepsilon^2 n^2}\Big)\,, \tag{8}$$

where we are interested in $\sigma^2 < 1$.

Figure 1: 2-d PCA under the Gaussian data from Remark 3.4 (left) shows that the average gradient (red arrow) is smaller than the range of the minibatch of 400 gradients (blue dots). Under Example 4.1 (right), the range can be made arbitrarily smaller than the average gradient.

Algorithm 3: Differentially Private Principal Component Analysis (DP-PCA)

Input: $S = \{A_i\}_{i=1}^n$, (ε, δ), batch size $B \in \mathbb{Z}_+$, learning rates $\{\eta_t\}_{t=1}^{\lfloor n/B \rfloor}$, probability $\zeta \in (0,1)$

Choose w_0 uniformly at random from the unit sphere

for $t = 1, 2, \ldots, T = \lfloor n/B \rfloor$ **do**

  Run Private Top Eigenvalue Estimation (Algorithm 4) with $(\varepsilon/2, \delta/2)$-DP and failure probability $\zeta/(2T)$ on $\{A_{B(t-1)+i} w_{t-1}\}_{i=1}^{\lfloor B/2 \rfloor}$. Let the returned estimation be $\hat{\Lambda}_t > 0$.

  Run Private Mean Estimation (Algorithm 5) with $(\varepsilon/2, \delta/2)$-DP, failure probability $\zeta/(2T)$, and the estimated eigenvalue $2\hat{\Lambda}_t$ on $\left\{A_{B(t-1)+\lfloor B/2 \rfloor + i} w_{t-1}\right\}_{i \in \lfloor B/2 \rfloor}$. Let the returned mean gradient estimate be $\hat{g}_t \in \mathbb{R}^d$.

  $w_t' \leftarrow w_{t-1} + \eta_t \hat{g}_t \quad , \qquad w_t \leftarrow w_t'/\|w_t'\|$

Return w_T

This is problematic since the second term, due to the DP noise, does not vanish as the randomness σ^2 in the data decreases. We do not observe this for Gaussian data where signal and noise scale proportionally as shown below. We reduce the noise we add for privacy, by switching from a simple norm clipping, that adds noise proportional to the norm of the gradients, to private estimation, that only requires the noise to scale as the *range* of the gradients, i.e. the maximum distance between two gradients in the minibatch. The toy example above showcases that the range can be arbitrarily smaller than the maximum norm (Fig. 1). We want to emphasize that although the idea of using private estimation within an optimization has been conceptually proposed in abstract settings, e.g., in [51], DP-PCA is the first setting where (i) such separation between the norm and the range of the gradients holds under any statistical model, and hence (ii) the long line of recent advances in private estimation provides significant gain over the simple DP-SGD [1].

5 Differentially Private Principal Component Analysis (DP-PCA)

Combining the two ideas of minibatch SGD and private mean estimation, we propose DP-SGD. We use minibatch SGD of minibatch size $B = O(n/\log^2 n)$ to allow for larger range of $\varepsilon = O(1)$. We use Private Mean Estimation to add an appropriate level of noise chosen adaptively according to Private Eigenvalue Estimation. We describe details of both sub-routines in Section 6.

We show an upper bound on the error achieved by DP-PCA under an appropriate choice of the learning rate. We provide a complete proof in Appendix E.1 that includes the explicit choice of the learning rate η_t in Eq. (67), and a proof sketch is provided in Section 6.1.

Theorem 5.1. *For $\varepsilon \in (0, 0.9)$, DP-PCA guarantees (ε, δ)-DP for all S, B, ζ, and δ. Given n i.i.d. samples $\{A_i \in \mathbb{R}^{d \times d}\}_{i=1}^n$ satisfying Assumption 1 with parameters $(\Sigma, M, V, K, \kappa, a, \gamma^2)$, if*

$$n \;=\; \tilde{O}\Big(e^{\kappa^2} + \frac{d^{1/2}(\log(1/\delta))^{3/2}}{\varepsilon} + \kappa M + \kappa^2 V + \frac{d\,\kappa\,\gamma\,(\log(1/\delta))^{1/2}}{\varepsilon} + \frac{d\log(1/\delta)}{\varepsilon} \Big)\,, \tag{9}$$

with a large enough constant and $\delta \leq 1/n$, then there exists a positive universal constant c_1 and a choice of learning rate η_t that depends on $(t, M, V, K, a, \lambda_1, \lambda_1 - \lambda_2, n, d, \varepsilon, \delta)$ such that $T = \lfloor n/B \rfloor$ steps of DP-PCA in Algorithm 3 with choices of $\zeta = 0.01$ and $B = c_1 n/(\log n)^2$, *outputs w_T such that with probability* 0.99*,*

$$\sin(w_T, v_1) \;=\; \widetilde{O}\left(\kappa\Big(\sqrt{\frac{V}{n}} + \frac{\gamma d\sqrt{\log(1/\delta)}}{\varepsilon n}\Big)\right)\;, \tag{10}$$

where $\widetilde{O}(\cdot)$ hides poly-logarithmic factors in n, d, $1/\varepsilon$, and $\log(1/\delta)$ and polynomial factors in K.

We further interpret this analysis and show that (i) DP-PCA is nearly optimal when the data is from a Gaussian distribution by comparing against a lower bound (Theorem 5.3); and (ii) DP-PCA significantly improves upon the private Oja's algorithm under Example 4.1. We discuss the necessity of some of the assumptions at the end of this section, including how to agnostically find the appropriate learning rate scheduling.

Near-optimality of DP-PCA under Gaussian distributions. Consider the case of i.i.d. samples $\{x_i\}_{i=1}^n$ from a Gaussian distribution from Remark 3.4.

Corollary 5.2 (Upper bound; Gaussian distribution)**.** *Under the hypotheses of Theorem 5.1 and $\{A_i = x_i x_i^\top\}_{i=1}^n$ with Gaussian random vectors x_i's, after $T = n/B$ steps, DP-PCA outputs w_T that achieves, with probability* 0.99*,*

$$\sin(w_T, v_1) \;=\; \tilde{O}\left(\kappa\left(\sqrt{\frac{d}{n}} + \frac{d\sqrt{\log(1/\delta)}}{\varepsilon n}\right)\right)\;. \tag{11}$$

We prove a nearly matching lower bound, up to factors of $\sqrt{\lambda_1/\lambda_2}$ and $\sqrt{\log(1/\delta)}$. One caveat is that the lower bound assumes *pure*-DP with $\delta = 0$. We do not yet have a lower bound technique for approximate DP that is tight, and all known approximate DP lower bounds have gaps to achievable upper bounds in its dependence in $\log(1/\delta)$, e.g., [5, 62]. We provide a proof in Appendix C.1.

Theorem 5.3 (Lower bound; Gaussian distribution)**.** *Let $\mathcal{M}_\varepsilon$ be a class of $(\varepsilon, 0)$-DP estimators that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of Gaussian distributions with (λ_1, λ_2) as the first and second eigenvalues of the covariance matrix is denoted by $\mathcal{P}_{(\lambda_1,\lambda_2)}$. For d>c where $c > 0$ is some absolute constant, there exists a universal constant $C > 0$ such that*

$$\inf_{\hat{v}\in\mathcal{M}_\varepsilon} \sup_{P\in\mathcal{P}_{(\lambda_1,\lambda_2)}} \mathbb{E}_{S\sim P^n}\left[\sin(\hat{v}(S), v_1)\right] \;\geq\; C\min\left(\kappa\left(\sqrt{\frac{d}{n}} + \frac{d}{\varepsilon n}\right)\sqrt{\frac{\lambda_2}{\lambda_1}}, 1\right)\;. \tag{12}$$

Comparisons with private Oja's algorithm. We demonstrate that DP-PCA can significantly improve upon Private Oja's Algorithm with Example 4.1, where DP-PCA achieves an error bound of $\sin(w_T, v_1) = \tilde{O}\big(\sigma\sqrt{d/n} + \sigma d\sqrt{\log(1/\delta)}/(\varepsilon n)\big)$. As the noise power σ^2 decreases DP-PCA achieves a vanishing error, whereas Private Oja's Algorithm has a non-vanishing error in Eq. (8). This follows from the fact that the second term in the error bound in Eq. (10) scales as γ, which can be made arbitrarily smaller than the second term in Eq. (5) that scales as $(\gamma + 1)$. Further, the error bound for DP-PCA holds for any $\varepsilon = O(1)$, whereas Private Oja's Algorithm requires significantly smaller $\varepsilon = O(\sqrt{\log(n/\delta)/n})$.

Remarks on the assumptions of Theorem 5.1. We have an exponential dependence of the sample complexity in the spectral gap, $n \geq \exp(\kappa^2)$. This ensures we have a large enough $T = \lfloor n/B \rfloor$ to reduce the non-dominant second term in Eq. (2), in balancing the learning rate η_t and T (which is explicitly shown in Eqs. 69 and (70) in the Appendix). It is possible to get rid of this exponential dependence at the cost of an extra term of $\tilde{O}(\kappa^4\gamma^2 d^2 \log(1/\delta)/(\varepsilon n)^2)$ in the error rate in Eq. (10), by selecting a slightly larger $T = c\kappa^2 \log^2 n$. A Gaussian-like tail bound in Assumption A.4 is necessary to get the desired upper bound scaling as $\tilde{O}(d\sqrt{\log(1/\delta)}/(\varepsilon n))$ in Eq. 11, for example. The next lower bound shows that without such assumptions on the tail, the error due to privacy scales as $\Omega(\sqrt{d \wedge \log(1/\delta)/(\varepsilon n)})$. We believe that the dependence in δ is loose, and it might be possible to get a tighter lower bound using [52]. We provide a proof and other lower bounds in Appendix C.

Theorem 5.4 (Lower bound without Assumption A.4)**.** *Let $\mathcal{M}_\varepsilon$ be a class of (ε, δ)-DP estimators that map n i.i.d. samples to an estimate $\hat{v} \in \mathbb{R}^d$. A set of distributions satisfying Assumptions A.1–A.3* *with $M = \tilde{O}(d + \sqrt{n\varepsilon/d})$, $V = O(d)$ and $\gamma = O(1)$ is denoted by $\tilde{\mathcal{P}}$. For $d \geq 2$, there exists a universal constant $C > 0$ such that*

$$\inf_{\hat{v} \in \mathcal{M}_\varepsilon} \sup_{P \in \tilde{\mathcal{P}}} \mathbb{E}_{S \sim P^n} \left[\sin(\hat{v}(S), v_1)\right] \; \geq \; C\kappa \min\left(\sqrt{\frac{d \wedge \log\left(\left(1 - e^{-\varepsilon}\right)/\delta\right)}{\varepsilon n}}, 1 \right) . \tag{13}$$

Currently, DP-PCA requires choices of the learning rates, η_t, that depend on possibly unknown quantities. Since we can privately evaluate the quality of our solution, one can instead run multiple instances of DP-PCA with varying $\eta_t = c_1/(c_2 + t)$ and find the best choice of $c_1 > 0$ and $c_2 > 0$. Let $w_T(c_1, c_2)$ denote the resulting solution for one instance of $\{\eta_t = c_1/(c_2 + t)\}_{t=1}^T$. We first set a target error ζ. For each round $i = 1, \ldots$, we will run algorithm for $(c_1, c_2) = [2^{i-1}, 2^{-i+1}] \times [2^{-i+1}, 2^{-i+2} \ldots, 2^{i-1}]$ and $(c_1, c_2) = [2^{-i+1}, 2^{-i+2} \ldots, 2^{i-1}] \times [2^{i-1}, 2^{-i+1}]$, and compute each $\sin(w_T(c_1, c_2), v_1)$ privately, each with privacy budget $\varepsilon_i = \frac{\varepsilon}{2^{i+1}(2i-1)}, \delta_i = \frac{\delta}{2^{i+1}(2i-1)}$. We terminate the algorithm once there there is a $w_T(c_1, c_2)$ satisfies $\sin(w_T(c_1, c_2), v_1) \leq \zeta$. It is clear that this search meta-algorithm terminate in logarithmic round, and the total sample complexity only blows up by a poly-log factor.

6 Private mean estimation for the minibatch stochastic gradients

DP-PCA critically relies on private mean estimation to reduce variance of the noise required to achieve (ε, δ)-DP. We follow a common recipe from [56, 50, 54, 8, 15]. First, we privately find an approximate range of the gradients in the minibatch (Alg. 4). Next, we apply the Gaussian mechanism to the truncated gradients where the truncation is tailored to the estimated range (Alg. 5).

Step 1: estimating the range. We need to find an approximate range of the minibatch of gradients in order to adaptively truncate the gradients and bound the sensitivity. Inspired by a private preconditioning mechanism designed for mean estimation with unknown covariance from [53], we propose to use privately estimated top eigenvalue of the covariance matrix of the gradients. For details on the version of the histogram learner we use in Alg. 4 in Appendix E.2, we refer to [61, Lemma D.1]. Unlike the private preconditioning of [53] that estimates all eigenvalues and requires $n = \widetilde{O}(d^{3/2}\log(1/\delta)/\varepsilon)$ samples, we only require the top eigenvalue and hence the next theorem shows that we only need $n = \widetilde{O}(d\log(1/\delta)/\varepsilon)$.

Theorem 6.1. *Algorithm 4 is (ε, δ)-DP. Let $g_i = A_i u$ for some fixed vector u, where A_i satisfies A.1 and A.4 in Assumption 1 such that the mean is $\mathbb{E}[g_i] = \Sigma u$ and the covariance is $\mathbb{E}[(g_i - \Sigma u)(g_i - \Sigma u)^\top] = \lambda_1^2 H_u$. With a large enough sample size scaling as*

$$B = O\left(\frac{K^2 \, d \, \log(d\log(1/(\delta\zeta))/(\zeta\varepsilon)) \log^{2a}(Bd/\zeta)\log(1/(\zeta\delta))}{\varepsilon}\right) = \tilde{O}\left(\frac{K^2 \, d \, \log(1/\delta)}{\varepsilon}\right),$$

Algorithm 4 outputs $\hat{\Lambda}$ achieving $\hat{\Lambda} \in \left[(1/\sqrt{2})\lambda_1^2\|H_u\|_2, \sqrt{2}\lambda_1^2\|H_u\|_2\right]$ with probability $1-\zeta$, where the pair $(K > 0, a > 0)$ parametrizes the tail of the distribution in A.4 and $\tilde{O}(\cdot)$ hides logarithmic factors in $B, d, 1/\zeta, \log(1/\delta)$, and ε.

We provide a proof in Appendix E.2. There are other ways to privately estimate the range. Some approaches require known bounds such as $\sigma_{\min}^2 \leq \lambda_1^2(H_u)_{ii} \leq \sigma_{\max}^2$ for all $i \in [d]$ [56], and other agnostic approaches are more involved such as instance optimal universal estimators of [17].

Step 2: Gaussian mechanism for mean estimation. Once we have a good estimate of the top eigenvalue from the previous section, we use it to select the bin size of the private histogram and compute the truncated empirical mean. Since truncated empirical mean has a bounded sensitivity, we can use Gaussian mechanism to achieve DP. The algorithm is now standard in DP mean estimation, e.g., [56, 50]. However, the analysis is slightly different since our assumptions on g_i's are different. For completeness, we provide the Algorithm 5 in Appendix E.3.

The next lemma shows that the Private Mean Estimation is (ε, δ)-DP, and with high probability clipping does not apply to any of the gradients. The returned private mean, therefore, is distributed as a spherical Gaussian centered at the empirical mean of the gradients. This result requires that we have a good estimate of the top eigenvalue from Alg. 4 such that $\hat{\Lambda} \simeq \lambda_1^2\|H_u\|_2$. This analysis implies that we get an unbiased estimate of the gradient mean (which is critical in the analysis) with noise scaling as $\tilde{O}(\lambda_1\gamma\sqrt{d\log(1/\delta)}/(\varepsilon B))$, where $\gamma^2 = \max_{u:\|u\|=1}\|H_u\|_2$ (which is critical in getting the tight sample complexity in the second term of the final utility guarantee in Eq. (10)). We provide a proof in Appendix E.3.

Lemma 6.2. *For $\varepsilon \in (0, 0.9)$ and any $\delta \in (0,1)$, Algorithm 5 is (ε,δ)-DP. Let $g_i = A_i u$ for some fixed vector u, where A_i satisfies A.1 and A.4 in Assumption 1 such that the mean is $\mathbb{E}[g_i] = \Sigma u$ and the covariance is $\mathbb{E}[(g_i - \Sigma u)(g_i - \Sigma u)^\top] = \lambda_1^2 H_u$. If $\hat{\Lambda} \in [\lambda_1^2\|H_u\|_2/\sqrt{2}, \sqrt{2}\lambda_1^2\|H_u\|_2]$, $\delta \leq 1/B$, and $B = \Omega((\sqrt{d\log(1/\delta)}/\varepsilon)\log(d/(\zeta\delta)))$ then, with probability $1-\zeta$, $g_i \in \bar{g} + \left[-3K\sqrt{\hat{\Lambda}}\log^a(Bd/\zeta), 3K\sqrt{\hat{\Lambda}}\log^a(Bd/\zeta)\right]^d$ for all $i \in [B]$.*

6.1 Proof sketch of Theorem 5.1

We choose $B = \Theta(n/\log^2 n)$ such that we access the dataset only $T = \Theta(\log^2 n)$ times. Hence we do not need to rely on amplification by shuffling. To add Gaussian noise that scales as the standard deviation of the gradients in each minibatch (as opposed to potentially excessively large mean of the gradients), DP-PCA adopts techniques from recent advances in private mean estimation. Namely, we first get a private and accurate estimate of the range from Theorem 6.1. Using this estimate, $\hat{\Lambda}$, Private Mean Estimation returns an unbiased estimate of the empirical mean of the gradients, as long as no truncation has been applied as ensured by Lemma 6.2. This gives

$$w_t' \leftarrow w_{t-1} + \eta_t\left(\frac{1}{B}\sum_{i=1}^{B} A_{B(t-1)+i}w_{t-1} + \beta_t z_t\right)\ , \tag{14}$$

for $z_t \sim \mathcal{N}(0,\mathbf{I})$ and $\beta_t = \frac{8K\sqrt{2\hat{\Lambda}_t}\log^a(Bd/\zeta)\sqrt{2d\log(2.5/\delta)}}{\varepsilon B}$. Using rotation invariance of spherical Gaussian random vectors and the fact that $\|w_{t-1}\| = 1$, we can reformulate it as

$$w_t' \leftarrow w_{t-1} + \eta_t\underbrace{\left(\frac{1}{B}\sum_{i=1}^{B} A_{B(t-1)+i} + \beta_t G_t\right)}_{\tilde{A}_t} w_{t-1}\ . \tag{15}$$

This process can be analyzed with Theorem 2.2 with $\tilde{A}_t$ substituting A_t.

7 Discussion

Under the canonical task of computing the principal component from i.i.d. samples, we show the first result achieving an optimal error rate. This critically relies on two ideas: minibatch SGD and private mean estimation. In particular, private mean estimation plays a critical role in the case when the range of the gradients is significantly smaller than the norm; we achieve an optimal error rate that cannot be achieved with the standard recipe of gradient clipping.

Assumption A.4 can be relaxed to heavy-tail bounds with bounded k-th moment on A_i, in which case we expect the second term in Eq. (10) to scale as $O(d(\sqrt{\log(1/\delta)}/\varepsilon n)^{1-1/k})$, drawing analogy from a similar trend in a computationally inefficient DP-PCA without spectral gap [62, Corollary 6.10]. When a fraction of data is corrupted, recent advances in [84, 58, 46] provide optimal algorithms for PCA. However, existing approach of [62] for robust and private PCA is computationally intractable. Borrowing ideas from robust and private mean estimation in [61], one can design an efficient algorithm, but at the cost of sub-optimal sample complexity. It is an interesting direction to design an optimal and robust version of DP-PCA. Our lower bounds are loose in its dependence in $\log(1/\delta)$. Recently, a promising lower bound technique has been introduced in [52] that might close this gap.

There are two ways to extend our framework to general rank-r PCA, whose analyses are promising research directions. First, applying Hotelling's deflation method [40], we can iteratively find the PCA components one by one, by alternating our DP-PCA and deflation. For example, in one step of the iteration, we only update the current iterate vector in the directions orthogonal to all the previously found PCA components. Repeating this steps gives the estimates of the top principal components. Secondly, we can directly apply Oja's algorithm. We keep track of a r-dimensional subspace in the Oja's update rule for PCA, and perform QR decomposition to keep the iterates on the Grassmannian manifold. It might be possible to extend the analysis of [42] to analyze the private version.

Acknowledgments and Disclosure of Funding

This work is supported in part by Google faculty research award and NSF grants CCF-1705007, CNS-2002664, IIS-1929955, DMS-2134012, CCF-2019844 as a part of NSF Institute for Foundations of Machine Learning (IFML), CNS-2112471 as a part of NSF AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE).

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] John M Abowd. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867, 2018.

[3] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

[4] Maria-Florina Balcan, Simon Shaolei Du, Yining Wang, and Adams Wei Yu. An improved gap-dependency analysis of the noisy power method. In *Conference on Learning Theory*, pages 284–309. PMLR, 2016.

[5] Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

[6] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 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, 2014.

[8] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *arXiv preprint arXiv:2006.06618*, 2020.

[9] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: the sulq framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 128–138, 2005.

[10] Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[11] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. *J. Mach. Learn. Res.*, 20:94–1, 2019.

[12] Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

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

[15] Christian Covington, Xi He, James Honaker, and Gautam Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*, 2021.

[16] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 291–305. Springer, 2014.

[17] Wei Dong and Ke Yi. Universal private estimators. *arXiv preprint arXiv:2111.02598*, 2021.

[18] John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In *Conference on Learning Theory*, pages 1161–1191. PMLR, 2019.

[19] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[20] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[21] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[22] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[23] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. 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, 2014.

[24] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[25] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[26] Hossein Esfandiari, Vahab Mirrokni, and Shyam Narayanan. Tight and robust private mean estimation with few users. *arXiv preprint arXiv:2110.11876*, 2021.

[27] Giulia Fanti, Vasyl Pihur, and Úlfar Erlingsson. Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, 2016(3):41–61, 2016.

[28] Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 439–449, 2020.

[29] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964. IEEE, 2022.

[30] Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pages 535–544. PMLR, 2018.

[31] James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

[32] Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation: z-test and tight confidence intervals. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2545–2554. PMLR, 2019.

[33] Arpita Gang, Bingqing Xiang, and Waheed U Bajwa. Distributed principal subspace analysis for partitioned big data: Algorithms, analysis, and implementation. *IEEE Transactions on Signal and Information Processing over Networks*, 7:699–715, 2021.

[34] Quan Geng, Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The staircase mechanism in differential privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1176–1184, 2015.

[35] Quan Geng and Pramod Viswanath. The optimal mechanism in differential privacy. In *2014 IEEE international symposium on information theory*, pages 2371–2375. IEEE, 2014.

[36] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.

[37] Moritz Hardt and Aaron Roth. Beating randomized response on incoherent matrices. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*, pages 1255–1268, 2012.

[38] Moritz Hardt and Aaron Roth. 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, 2013.

[39] Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. *arXiv preprint arXiv:2111.12981*, 2021.

[40] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[41] Lijie Hu, Shuo Ni, Hanshen Xiao, and Di Wang. High dimensional differentially private stochastic optimization with heavy-tailed data. *arXiv preprint arXiv:2107.11136*, 2021.

[42] De Huang, Jonathan Niles-Weed, and Rachel Ward. Streaming k-pca: Efficient guarantees for oja's algorithm, beyond rank-one updates. In *Conference on Learning Theory*, pages 2463–2498. PMLR, 2021.

[43] Hafiz Imtiaz and Anand D Sarwate. Differentially private distributed principal component analysis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2206–2210. IEEE, 2018.

[44] Hafiz Imtiaz and Anand D Sarwate. Distributed differentially private algorithms for matrix and tensor factorization. *IEEE journal of selected topics in signal processing*, 12(6):1449–1464, 2018.

[45] 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. In *Conference on learning theory*, pages 1147–1164. PMLR, 2016.

[46] Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems*, 33, 2020.

[47] Matthew Joseph, Janardhan Kulkarni, Jieming Mao, and Steven Z Wu. Locally private gaussian estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

[48] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, 27, 2014.

[49] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[50] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. PMLR, 2019.

[51] Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. *arXiv preprint arXiv:2106.01336*, 2021.

[52] Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. *arXiv preprint arXiv:2205.08532*, 2022.

[53] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. *arXiv preprint arXiv:2111.04609*, 2021.

[54] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. *arXiv preprint arXiv:2002.09464*, 2020.

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

[56] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

[57] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

[58] Weihao Kong, Raghav Somani, Sham Kakade, and Sewoong Oh. Robust meta-learning for mixed linear regression with small batches. *Advances in Neural Information Processing Systems*, 33, 2020.

[59] Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. *arXiv preprint arXiv:2112.03548*, 2021.

[60] Janardhan Kulkarni, Yin Tat Lee, and Daogao Liu. Private non-smooth empirical risk minimization and stochastic convex optimization in subquadratic steps. *arXiv preprint arXiv:2103.15352*, 2021.

[61] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34, 2021.

[62] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. PMLR, 2022.

[63] Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[64] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[65] Frank D McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

[66] Jason Milionis, Alkis Kalavasis, Dimitris Fotakis, and Stratis Ioannidis. Differentially private regression with unbounded covariates. In *International Conference on Artificial Intelligence and Statistics*, pages 3242–3273. PMLR, 2022.

[67] Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, pages 956–964, 2016.

[68] Darakhshan J Mir. *Differential privacy: an exploration of the privacy-utility landscape*. Rutgers The State University of New Jersey-New Brunswick, 2013.

[69] Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

[70] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes for course 18S997*, 813(814):46, 2015.

[71] Or Sheffet. Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pages 789–827. PMLR, 2019.

[72] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

[73] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[74] Christos Tzamos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Ilias Zadik. Optimal private median estimation under minimal distributional assumptions. *Advances in Neural Information Processing Systems*, 33:3301–3311, 2020.

[75] Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression via adaptive clipping. *arXiv preprint arXiv:2207.04686*, 2022.

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

[77] Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

[78] Vincent Vu and Jing Lei. Minimax rates of estimation for sparse pca in high dimensions. In *Artificial intelligence and statistics*, pages 1278–1286. PMLR, 2012.

[79] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[80] Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *International Conference on Machine Learning*, pages 10081–10091. PMLR, 2020.

[81] Sen Wang and J Morris Chang. Differentially private principal component analysis over horizontally partitioned data. In *2018 IEEE Conference on Dependable and Secure Computing (DSC)*, pages 1–8. IEEE, 2018.

[82] Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

[83] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.

[84] Huan Xu, Constantine Caramanis, and Shie Mannor. Principal component analysis with contaminated data: The high dimensional case. *arXiv preprint arXiv:1002.4658*, 2010.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , see Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is theoretical and does not have a direct negative societal impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] We don't have experiments
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A] We didn't use existing assets.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]