Appendices

Supplementary Material for "Near-optimal Differentially Private Principal Components"

This supplement contains the proofs of our results as well as additional details of the experiments and the implementation of the exponential mechanism. Citation numbers here refer to the bibliography at the end of the supplement and not those of the main document.

We will write $\mathbf{KL}\left(f\|g\right) = \int f(x)\frac{f(x)}{g(x)}dx$ for the Kullback-Leibler divergence between two densities f and g.

A The algorithms guarantee privacy

We first describe the simple proofs that MOD-SULQ and PPCA guarantee differential privacy.

A.1 Proof of Theorem 1

Proof of Theorem 1. Let B and $\hat{B}$ be two independent symmetric random matrices where $\{B_{ij} : 1 \leq i \leq j \leq d\}$ and $\{\hat{B}_{ij} : 1 \leq i \leq j \leq d\}$ are each sets of i.i.d. Gaussian random variables with mean 0 and variance β^2. Consider two data sets $\mathcal{D} = \{x_i : i = 1, 2, \ldots, n\}$ and $\hat{\mathcal{D}} = \mathcal{D}_1 \cup \{\hat{x}_n\} \setminus \{x_n\}$ and let A and $\hat{A}$ denote their second moment matrices. Let $G = A + B$ and $\hat{G} = \hat{A} + \hat{B}$. We first calculate the log ratio of the densities of G and $\hat{G}$ at a point H:

$$\begin{aligned}\log\frac{f_G(H)}{f_{\hat{G}}(H)} &= \sum_{1\leq i\leq j\leq d}\left(-\frac{1}{2\beta^2}(H_{ij}-A_{ij})^2 + \frac{1}{2\beta^2}(H_{ij}-\hat{A}_{ij})^2\right)\\ &= \frac{1}{2\beta^2}\sum_{1\leq i\leq j\leq d}\left(\frac{2}{n}(H_{ij}-A_{ij})(x_{n,i}x_{n,j}-\hat{x}_{n,i}\hat{x}_{n,j}) + \frac{1}{n^2}(\hat{x}_{n,i}\hat{x}_{n,j}-x_{n,i}x_{n,j})^2\right).\end{aligned}$$

From (16) the last term is upper bounded by $2/n^2$. To upper bound the first term,

$$\begin{aligned}\sum_{1\leq i\leq j\leq d}|\hat{x}_{n,i}\hat{x}_{n,j}-x_{n,i}x_{n,j}| &\leq 2\max_{a:\|a\|\leq 1}\sum_{1\leq i\leq j\leq d}a_ia_j\\ &\leq 2\cdot\frac{1}{2}(d^2+d)\cdot\frac{1}{d}\\ &= d+1.\end{aligned}$$

Note that this bound is not too loose – by taking $\hat{x} = d^{-1/2}\mathbf{1}$ and $x = (1, 0, \ldots, 0)^T$, this term is still linear in d.

Then for any measurable set $\mathcal{S}$ of matrices,

$$\mathbb{P}\left(G\in\mathcal{S}\right) \leq \exp\left(\frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\gamma + \frac{3}{n^2}\right)\right)\mathbb{P}\left(\hat{G}\in\mathcal{S}\right) + \mathbb{P}\left(B_{ij} > \gamma \text{ for all } i,j\right). \quad (9)$$

To handle the last term, use a union bound over the $(d^2+d)/2$ variables $\{B_{ij}\}$ together with the tail bound, which holds for $\gamma > \beta$:

$$\mathbb{P}\left(B_{ij} > \gamma\right) \leq \frac{1}{\sqrt{2\pi}}e^{-\gamma^2/2\beta^2}.$$

Thus setting $\mathbb{P}\left(B_{ij} > \gamma \text{ for some } i,j\right) = \delta$ yields the condition

$$\delta = \frac{d^2+d}{2\sqrt{2\pi}}e^{-\gamma^2/2\beta^2}.$$

Rearranging to solve for γ gives

$$\gamma = \max\left(\beta, \beta\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)}\right) = \beta\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)}$$

for $d > 1$ and $\delta < 3/\sqrt{2\pi e}$. This then gives an expression for α to make (9) imply (α, δ) differential privacy:

$$\begin{aligned}\alpha &= \frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\gamma + \frac{2}{n^2}\right) \\ &= \frac{1}{2\beta^2}\left(\frac{2}{n}(d+1)\beta\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{2}{n^2}\right).\end{aligned}$$

Solving for β using the quadratic formula yields a particularly messy expression:

$$\begin{aligned}\beta &= \frac{d+1}{2n\alpha}\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{1}{2n\alpha}\left(2(d+1)^2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right) + 4\alpha\right)^{1/2} \\ &\leq \frac{d+1}{n\alpha}\sqrt{2\log\left(\frac{d^2+d}{\delta 2\sqrt{2\pi}}\right)} + \frac{1}{\sqrt{\alpha}n}. \qquad (10)\end{aligned}$$

□

A.2 Proof of Theorem 2

Proof of Theorem 2. Let X be a data matrix whose i-th column is x_i and $A = \frac{1}{n}XX^T$. The PP-PCA algorithm is the exponential mechanism of McSherry and Talwar [28] applied to the score function

$$q_{\mathrm{F}}(X, v) = n \cdot v^T A v$$

Consider $X' = [x_1\ x_2\ \cdots\ x_{n-1}\ x'_n]$ differ from X in a single column and let $A' = \frac{1}{n}X'X'^T$. We have

$$\begin{aligned}\max_{v\in\mathbb{S}^{d-1}} |q_{\mathrm{F}}(X', v) - q_{\mathrm{F}}(X, v)| &\leq \left|v^T(x'_n x'^T_n - x_n x_n^T)v\right| \\ &\leq \left|\left\|v^T x'_n\right\|^2 - \left\|v^T x_n\right\|^2\right| \\ &\leq 1.\end{aligned}$$

The last step follows because $\|x_i\| \leq 1$ for all i. The result now follows immediately from the results of McSherry and Talwar [28, Theorem 6]. □

B Proof of Theorem 3

The results on the exponential mechanism [28] bound the gap between the value of the function $q_{\mathrm{F}}(\hat{v}_1) = n\cdot\hat{v}_1^T A\hat{v}_1$ evaluated at the output $\hat{v}_1$ of the mechanism and the optimal value $q(v_1) = n\cdot\lambda_1$. We derive a bound on the correlation $q_{\mathrm{A}}(\hat{v}_1) = |\langle\hat{v}_1, v_1\rangle|$ via geometric arguments.

Lemma 1 (Lemmas 2.2 and 2.3 of Ball [3])**.** *Let μ be the uniform measure on the unit sphere $\mathbb{S}^{d-1}$. For any $x \in \mathbb{S}^{d-1}$ and $0 \leq c < 1$*

$$\frac{1}{2}\exp\left(-\frac{d-1}{2}\log\frac{2}{1-c}\right) \leq \mu\left(\left\{v \in \mathbb{S}^{d-1} : \langle v, x\rangle \geq c\right\}\right) \leq \exp\left(-dc^2/2\right). \qquad (11)$$

Proof of Theorem 3. Fix a privacy level α, target correlation ρ, and probability η. Let X be the data matrix and $B = (\alpha/2)XX^T$ and

$$\mathcal{U}_\rho = \{u : |\langle u, v_1\rangle| \geq \rho\}.$$

be the union of the two spherical caps centered at $\pm v_1$. Let $\overline{\mathcal{U}}_\rho$ denote the complement of $\mathcal{U}_\rho$ in $\mathbb{S}^{d-1}$.

An output vector $\hat{v}_1$ is "good" if it is in $\mathcal{U}_\rho$. We first give some bounds on the score function $q_{\mathrm{F}}(u)$ on the boundary between $\mathcal{U}_\rho$ and $\overline{\mathcal{U}}_\rho$, where $\langle u, v_1\rangle = \pm\rho$. The function $q_{\mathrm{F}}(u)$ is maximized when u is a linear combination of v_1 and v_2, the top two eigenvectors of A. It minimized when u is a linear combination of v_1 and v_d. Therefore

$$q_{\mathrm{F}}(u) \le \frac{n\alpha}{2}(\rho^2\lambda_1 + (1-\rho^2)\lambda_2) \qquad u \in \overline{\mathcal{U}}_\rho \tag{12}$$

$$q_{\mathrm{F}}(u) \ge \frac{n\alpha}{2}(\rho^2\lambda_1 + (1-\rho^2)\lambda_d) \qquad u \in \mathcal{U}_\rho. \tag{13}$$

Let $\mu(\cdot)$ denote the uniform measure on the unit sphere. Then fixing an $0 \le b < 1$, using (12), (13), and the fact that $\lambda_d \ge 0$,

$$\begin{aligned}
\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right) &\le \frac{\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right)}{\mathbb{P}\left(\mathcal{U}_\sigma\right)} \\
&= \frac{\frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}m, B\right)} \int_{\overline{\mathcal{U}}_\rho} \exp\left(u^T B u\right) d\mu}{\frac{1}{{}_1F_1\left(\frac{1}{2}k, \frac{1}{2}m, B\right)} \int_{\mathcal{U}_\sigma} \exp\left(u^T B u\right) d\mu} \\
&\le \frac{\exp\left(n(\alpha/2)\left(\rho^2\lambda_1 + (1-\rho^2)\lambda_2\right)\right) \cdot \mu\left(\overline{\mathcal{U}}_\rho\right)}{\exp\left(n(\alpha/2)\left(\sigma^2\lambda_1 + (1-\sigma^2)\lambda_d\right)\right) \cdot \mu\left(\mathcal{U}_\sigma\right)} \\
&\le \exp\left(-\frac{n\alpha}{2}\left(\sigma^2\lambda_1 - (\rho^2\lambda_1 + (1-\rho^2)\lambda_2)\right)\right) \cdot \frac{\mu\left(\overline{\mathcal{U}}_\rho\right)}{\mu\left(\mathcal{U}_\sigma\right)}.
\end{aligned} \tag{14}$$

Applying the lower bound from Lemma 1 to the denominator of (14) and the upper bound $\mu\left(\overline{\mathcal{U}}_\rho\right) \le 1$ yields

$$\mathbb{P}\left(\overline{\mathcal{U}}_\rho\right) \le \exp\left(-\frac{n\alpha}{2}\left(\sigma^2\lambda_1 - (\rho^2\lambda_1 + (1-\rho^2)\lambda_2)\right)\right) \cdot \exp\left(\frac{d-1}{2}\log\frac{2}{1-\sigma}\right). \tag{15}$$

We must choose a $\sigma^2 > \rho^2$ to make the upper bound ≤ 1, but more precisely,

$$\sigma^2 > \rho^2 + (1-\rho^2)\frac{\lambda_2}{\lambda_1}$$

$$1 - \sigma^2 < (1-\rho^2)\left(1 - \frac{\lambda_2}{\lambda_1}\right)$$

For simplicity, choose

$$1 - \sigma^2 = \frac{1}{2}(1-\rho^2)\left(1 - \frac{\lambda_2}{\lambda_1}\right).$$

So that

$$\begin{aligned}
\sigma^2\lambda_1 - (\rho^2\lambda_1 + (1-\rho^2)\lambda_2) &= (1-\rho^2)\lambda_1 - (1-\sigma^2)\lambda_1 - (1-\rho^2)\lambda_2 \\
&= (1-\rho^2)\left(\lambda_1 - \frac{1}{2}(\lambda_1 - \lambda_2) - \lambda_2\right) \\
&= \frac{1}{2}(1-\rho^2)(\lambda_1 - \lambda_2),
\end{aligned}$$

and

$$\begin{aligned}
\log\frac{2}{1-\sigma} &< \log\frac{2}{1-\sigma^2} \\
&= \log\frac{4\lambda_1}{(1-\rho^2)(\lambda_1 - \lambda_2)}.
\end{aligned}$$

Setting the right hand side of (15) less than or equal to η yields

$$\frac{n\alpha}{4}(1-\rho^2)(\lambda_1 - \lambda_2) > \log\frac{1}{\eta} + \frac{d-1}{2}\log\frac{4\lambda_1}{(1-\rho^2)(\lambda_1 - \lambda_2)}.$$

Since $1-\rho < 1-\rho^2$, if we choose

$$n > \frac{d}{\alpha(1-\rho)(\lambda_1-\lambda_2)}\left(\frac{\log(1/\eta)}{d} + \log\frac{4\lambda_1}{(1-\rho^2)(\lambda_1-\lambda_2)}\right),$$

then the output of PPCA will produce a $\hat{v}_1$ such that

$$\mathbb{P}\left(|\langle \hat{v}_1, v_1\rangle| < \rho\right) \leq \eta.$$

□

In general, it is difficult to measure the area on the unit sphere of the set $\{x : x^T A x \geq 1-\gamma\}$. In the case where $A = I$ this is just a spherical cap, but for general A it can have a more irregular shape. The second reason is that explicit bounds on the confluent hypergeometric function with matrix argument do not give clear dependencies on the problem parameters.

C Proof of Theorem 5

In this section we provide theoretical guarantees on the performance of the MOD-SULQ algorithm. Theorem 1 shows that MOD-SULQ is (α, δ)-differentially private. Theorem 7 provides a lower bound on the distance between the vector released by MOD-SULQ and the true top eigenvector in terms of the privacy parameters α and δ and the number of points n in the data set. This implicitly gives a lower bound on the sample complexity of MOD-SULQ. We provide some graphical illustration of this tradeoff.

The following upper bound will be useful for future calculations : for two unit vectors x and y,

$$\sum_{1\leq i\leq j\leq d} (x_i x_j - y_i y_j)^2 \leq 2. \tag{16}$$

Note that this upper bound is achievable by setting x and y to be orthogonal elementary vectors.

The main tool in our lower bound is a generalization by Yu [37] of an information-theoretic inequality due to Fano.

Theorem 6 (Fano's inequality [37])**.** *Let $\mathcal{R}$ be a set and Θ be a parameter space with a pseudometric $d(\cdot)$. Let $\mathcal{F}$ be a set of r densities $\{f_1, \ldots, f_r\}$ on $\mathcal{R}$ corresponding to parameter values $\{\theta_1, \ldots, \theta_r\}$ in Θ. Let X have distribution $f \in \mathcal{F}$ with corresponding parameter θ and let $\hat{\theta}(X)$ be an estimate of θ. If, for all i and j*

$$d(\theta_i, \theta_j) \geq \tau \tag{17}$$

and

$$\mathbf{KL}\left(f_i \| f_j\right) \leq \gamma, \tag{18}$$

then

$$\max_j \mathbb{E}_j[d(\hat{\theta}, \theta_j)] \geq \frac{\tau}{2}\left(1 - \frac{\gamma + \log 2}{\log r}\right), \tag{19}$$

where $\mathbb{E}_j[\cdot]$ denotes the expectation with respect to distribution f_j.

To use this inequality, we will construct a set of densities on the set of covariance matrices corresponding distribution of the random matrix in the MOD-SULQ algorithm under different inputs. These inputs will be chosen using a set of unit vectors which are a packing on the surface of the unit sphere.

C.1 A packing lemma

The proof of this lemma is relatively straightforward. The following is a slight refinement of a lemma due to Csiszár and Narayan [11, 12].

Lemma 2. *Let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$ be arbitrary random variables and let $f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)$ be arbitrary with $0 \leq f_i \leq 1$, $i = 1, 2, \ldots, N$. Then the condition*

$$\mathbb{E}\left[f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) | \mathbf{Z}_1, \ldots, \mathbf{Z}_{i-1}\right] \leq a_i \text{ a.s.}, \qquad i = 1, 2, \ldots, N \tag{20}$$

implies that

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \leq \exp\left(-Nt(\log 2) + \sum_{i=1}^{N} a_i\right). \tag{21}$$

Proof. First apply Markov's inequality:

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \\
&\quad = \mathbb{P}\left(2^{\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)} > 2^{Nt}\right) \\
&\quad \leq 2^{-Nt}\mathbb{E}\left[2^{\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)}\right] \\
&\quad \leq 2^{-Nt}\mathbb{E}\left[2^{\sum_{i=1}^{N-1} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)}\mathbb{E}\left[2^{f_N(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)} | \mathbf{Z}_1, \ldots, \mathbf{Z}_{N-1}\right]\right].
\end{aligned}$$

Now note that for $b \in [0, 1]$ we have $2^b \leq 1 + b$, so

$$\begin{aligned}
\mathbb{E}\left[2^{f_N(\mathbf{Z}_1, \ldots, \mathbf{Z}_N)} | \mathbf{Z}_1, \ldots, \mathbf{Z}_{N-1}\right] &\leq \mathbb{E}\left[1 + f_N(\mathbf{Z}_1, \ldots, \mathbf{Z}_N) | \mathbf{Z}_1, \ldots, \mathbf{Z}_{N-1}\right] \\
&\leq (1 + a_N) \\
&\leq \exp(a_N).
\end{aligned}$$

Therefore

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \leq \exp(-Nt(\log 2) + a_N)\mathbb{E}\left[2^{\sum_{i=1}^{N-1} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i)}\right].$$

Continuing in the same way yields

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \leq \exp\left(-Nt(\log 2) + \sum_{i=1}^{N} a_i\right).$$

□

The second technical lemma [12, Lemma 2] is a basic result about the distribution of inner product between a randomly chosen unit vector and any other fixed vector. It is a consequence of a result of Shannon [34] on the distribution of the angle between a uniformly distributed unit vector and a fixed unit vector.

Lemma 3 (Lemma 2 of [12])**.** *Let $\mathbf{U}$ be uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$. Then for every unit vector $\mathbf{u}$ on this sphere and any $\phi \in [(2\pi d)^{-1/2}, 1)$, the following inequality holds:*

$$\mathbb{P}\left(\langle \mathbf{U}, \mathbf{u}\rangle \geq \phi\right) \leq (1 - \phi^2)^{(d-1)/2}. \tag{22}$$

Lemma 4 (Packing set on the unit sphere)**.** *Let a dimension d and $\phi \in [(2\pi d)^{-1/2}, 1)$ be given. For N and t satisfying*

$$-Nt(\log 2) + N(N-1)(1 - \phi^2)^{(d-1)/2} < 1, \tag{23}$$

there exists a set of $K = \lfloor (1-t)N \rfloor$ unit vectors $\mathcal{C}$ such that for all distinct pairs $\mu, \nu \in \mathcal{C}$,

$$|\langle \mu, \nu\rangle| < \phi. \tag{24}$$

Proof. The goal is to generate a set of N unit vectors on the surface of the sphere $\mathbb{S}^{d-1}$ such that they have large pairwise distances, or correspondingly small pairwise inner products. To that end, define $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_N$ i.i.d. uniformly distributed on $\mathbb{S}^{d-1}$ and

$$f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) = \mathbf{1}\left(|\langle \mathbf{Z}_i, \mathbf{Z}_j \rangle| > \phi,\ j < i\right). \tag{25}$$

That is, $f_i = 1$ if $\mathbf{Z}_i$ has large inner product with any $\mathbf{Z}_j$ for $j < i$. The conditional expectation, by a union bound and Lemma 3, is

$$\mathbb{E}\left[f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) | \mathbf{Z}_1, \ldots, \mathbf{Z}_{i-1}\right] \leq 2(i-1)(1-\phi^2)^{(d-1)/2}. \tag{26}$$

Let $a_i = (i-1)(1-\phi^2)^{(d-1)/2}$. Then

$$\sum_{i=1}^{N} a_i = N(N-1)(1-\phi^2)^{(d-1)/2}. \tag{27}$$

Then Lemma 2 shows

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} f_i(\mathbf{Z}_1, \ldots, \mathbf{Z}_i) > t\right) \leq \exp\left(-Nt(\log 2) + N(N-1)(1-\alpha^2)^{(d-1)/2}\right). \tag{28}$$

This inequality implies that as long as

$$-Nt(\log 2) + N(N-1)(1-\phi^2)^{(d-1)/2} < 1, \tag{29}$$

then there is a finite probability that $\{\mathbf{Z}_i\}$ contains a subset $\{\mathbf{Z}'_i\}$ of size $\lfloor (1-t)N \rfloor$ such that $\left|\langle \mathbf{Z}'_i, \mathbf{Z}'_j \rangle\right| < \phi$ for all (i, j). Therefore such a set exists. □

A simple setting of the parameters gives the following packing.

Lemma 5 (Simple packing set)**.** *For $\phi \in [(2\pi d)^{-1/2}, 1)$, there exists a set of*

$$K = \frac{1}{8}\exp\left((d-1)\log\frac{1}{\sqrt{1-\phi^2}}\right) \tag{30}$$

vectors $\mathcal{C}$ in $\mathbb{S}^{d-1}$ such that for any pair $\mu, \nu \in \mathcal{C}$, the inner product between them satisfies

$$|\langle \mu, \nu \rangle| \leq \phi. \tag{31}$$

Proof. Applying Lemma 4 yields a set of K vectors $\mathcal{C}$ satisfying (23) and (24). To get a simple bound that's easy to work with, we can set

$$-Nt(\log 2) + N(N-1)(1-\phi^2)^{(d-1)/2} - 1 = 0, \tag{32}$$

and find an N close to this. Setting $\psi = (1-\phi^2)^{(d-1)/2}$, the quadratic formula solving for N yields

$$\begin{aligned} N &= \frac{1}{\psi}\left(t\log 2 + \psi + \left((t\log 2 + \psi)^2 + 4\psi\right)^{1/2}\right) \\ &> \frac{t}{2\psi}. \end{aligned}$$

Now setting $K = \frac{t(1-t)}{2\psi}$ and $t = 1/2$ gives (30). So there exists a set of K vectors on $\mathbb{S}^{d-1}$ whose pairwise inner products are smaller than ϕ. □

The maximum set of points that can be selected on a sphere of dimension d such that their pairwise inner products are bounded by ϕ is an open question. These sets are sometimes referred to as spherical codes [9] because they correspond to a set of signaling points of dimension d that can be perfectly decoded over a channel with bounded noise. The bounds here are from a probabilistic construction and can be tightened for smaller d. However, in terms of scaling with d this construction is essentially optimal [34].

C.2 Lower bounds for input perturbation

Lemma 6. *Let Σ be a positive definite matrix and let f denote the density $\mathcal{N}(a, \Sigma)$ and g denote the density $\mathcal{N}(b, \Sigma)$. Then* $\mathbf{KL}(f\|g) = \frac{1}{2}(a-b)^T\Sigma^{-1}(a-b)$.

Proof. This is a simple calculation:

$$\begin{aligned}\mathbf{KL}(f\|g) &= \mathbb{E}_{x\sim f}\left[-\frac{1}{2}(x-a)^T\Sigma^{-1}(x-a) + \frac{1}{2}(x-b)\Sigma^{-1}(x-b)\right] \\ &= \frac{1}{2}\left(a^T\Sigma^{-1}a - a^T\Sigma^{-1}b - b^T\Sigma^{-1}a + b^T\Sigma^{-1}b\right) \\ &= \frac{1}{2}(a-b)^T\Sigma^{-1}(a-b).\end{aligned}$$

□

The next theorem is a lower bound on the expected distance between the vector output by MOD-SULQ and the true top eigenvector. In order to get this lower bound, we construct a class of data sets and use Fano's inequality to derive a bound on the minimax error over the class.

Theorem 7 (Utility bound for MOD-SULQ)**.** *Let d, n, and $\alpha > 0$ be given and let β be given by Algorithm 1 so that the output of MOD-SULQ is (α, δ)-differentially private for all data sets in $\mathbb{R}^d$ with n elements. Then there exists a data set with n elements such that if $\hat{v}_1$ denotes the output of MOD-SULQ and v_1 is the top eigenvector of the empirical covariance matrix of the data set, the expected correlation $\langle \hat{v}_1, v_1\rangle$ is upper bounded:*

$$\mathbb{E}\left[|\langle \hat{v}_1, v_1\rangle|\right] \leq \min_{\phi\in\Phi}\left(1 - \frac{(1-\phi)}{4}\left(1 - \frac{1/\beta^2 + \log 2}{(d-1)\log\frac{1}{\sqrt{1-\phi^2}} - \log(8)}\right)^2\right) \tag{33}$$

where

$$\Phi \in \left[\max\left\{\frac{1}{\sqrt{2\pi d}}, \sqrt{1-\exp\left(-\frac{2\log(8d)}{d-1}\right)}, \sqrt{1-\exp\left(-\frac{2/\beta^2 + \log(256)}{d-1}\right)}\right\}, 1\right). \tag{34}$$

Proof. For $\phi \in [(2\pi d)^{-1/2}, 1)$, Lemma 5 shows there exists a set of K unit vectors $\mathcal{C}$ such that for $\mu, \nu \in \mathcal{C}$, the inner product between them satisfies $|\langle \mu, \nu\rangle| < \phi$, where K is given by (30). Note that for small ϕ this setting of K is loose, but any orthonormal basis provides d unit vectors which are orthogonal, setting $K = d$ and solving for ϕ yields

$$\left(1 - \exp\left(-\frac{2\log(8d)}{d-1}\right)\right)^{1/2}.$$

Setting the lower bound on ϕ to the maximum of these two yields the set of ϕ and K which we will consider in (34).

For any unit vector μ, let

$$A(\mu) = \mu\mu^T + N, \tag{35}$$

where N is a $d \times d$ symmetric random matrix such that $\{N_{ij} : 1 \leq i \leq j \leq d\}$ are i.i.d. $\mathcal{N}(0, \beta^2)$, where β^2 is the noise variance used in the MOD-SULQ algorithm. Due to symmetry, the matrix $A(\mu)$ can be thought of as a jointly Gaussian random vector on the $d(d+1)/2$ variables $\{A_{ij}(\mu) : 1 \leq i \leq j \leq d\}$. The mean of this vector is

$$\bar{\mu} = \left(\mu_1^2, \mu_2^2, \ldots, \mu_d^2, \mu_1\mu_2, \mu_1\mu_3, \ldots, \mu_{d-1}\mu_d\right)^T, \tag{36}$$

and the covariance is $\beta^2 I_{d(d+1)/2}$. Let f_μ denote the density of this vector.

For $\mu, \nu \in \mathcal{C}$, the divergence between f_μ and f_ν can be calculated using Lemma 6:

$$\begin{aligned}\mathbf{KL}\left(f_\mu \| f_\nu\right) &= \frac{1}{2}(\bar{\mu}-\bar{\nu})^T\Sigma^{-1}(\bar{\mu}-\bar{\nu}) \\ &= \frac{1}{2\beta^2}\left\|\bar{\mu}-\bar{\nu}\right\|^2 \\ &\leq \frac{1}{\beta^2}. \end{aligned} \tag{37}$$

The last line follows from the fact that the vectors in $\mathcal{C}$ are unit norm.

For any two vectors $\mu, \nu \in \mathcal{C}$, lower bound the Euclidean distance between them using the upper bound on the inner product:

$$\|\mu - \nu\| \geq \sqrt{2(1-\phi)}. \tag{38}$$

Let $\Theta = \mathbb{S}^{d-1}$ with the Euclidean norm and $\mathcal{R}$ be the set of distributions $\{A(\mu) : \mu \in \Theta\}$. From (38) and (37), the set $\mathcal{C}$ satisfies the conditions of Theorem 6 with $\mathcal{F} = \{f_\mu : \mu \in \mathcal{C}\}$, $r = K$, $\tau = \sqrt{2(1-\phi)}$, and $\gamma = \frac{1}{\beta^2}$. The conclusion of the Theorem shows that for MOD-SULQ,

$$\max_{\mu \in \mathcal{C}} \mathbb{E}_{f_\mu}\left[\|\hat{v} - \mu\|\right] \geq \frac{\sqrt{2(1-\phi)}}{2}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right). \tag{39}$$

This lower bound is vacuous when the term inside the parenthesis is negative, which imposes further conditions on ϕ. Setting $\log K = 1/\beta^2 + \log 2$, we can solve to find another lower bound on ϕ:

$$\phi \geq \sqrt{1 - \exp\left(-\frac{2/\beta^2 + \log(256)}{d-1}\right)}. \tag{40}$$

This yields the third term in (34). Note that for larger n this term will dominate the others.

Using Jensen's inequality on the the left side of (39):

$$\max_{\mu \in \mathcal{C}} \mathbb{E}_{f_\mu}\left[2(1 - |\langle \hat{v}, \mu \rangle|)\right] \geq \frac{(1-\phi)}{2}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right)^2.$$

So there exists a $\mu \in \mathcal{C}$ such that

$$\mathbb{E}_{f_\mu}\left[|\langle \hat{v}, \mu \rangle|\right] \leq 1 - \frac{(1-\phi)}{4}\left(1 - \frac{1/\beta^2 + \log 2}{\log K}\right)^2. \tag{41}$$

Consider the data set consisting of n copies of μ. The corresponding covariance matrix is $\mu\mu^T$ with top eigenvector $v_1 = \mu$. The output of the algorithm MOD-SULQ applied to this data set is an estimator of μ and hence satisfies (41). Minimizing over ϕ gives the desired bound. □

The minimization over ϕ in (33) does not lead to analytically pretty results, but numerical optimization can give some insight into these bounds. In all experiments we set $\delta = 0.01$. Figure 3 shows the lower bound on the correlation $\langle \hat{v}_1, v_1 \rangle$ for MOD-SULQ as a function of the dimension for four different values of α. For large data set sizes and large α, the lower bound is not very tight, so for 10^7 data points MOD-SULQ may not suffer much performance, even for large dimensions. However, for smaller α the bound becomes sharper, especially for smaller data sets. To see the dependence on α, Figure 4 shows the correlation as a function of α for smaller values of d. As f increases, MOD-SULQ requires more and more data to produce an output which is correlated with the true top eigenvector. For example, in dimension 1024, for $\alpha = 3$, if $n = 10000$ the expected inner product is is lower bounded by ≈ 0.87, which corresponds to an angle of $30°$. Finally, Figure 5 shows the correlation as a function of n for different dimensions and different values of α. Again, in high dimension, the lower bound is shows that the expected performance of MOD-SULQ is poor when there are a small number of data points. This limitation may be particularly acute when the data lies in a very low dimensional subspace but is presented in very high dimension. In such "sparse" settings, perturbing the input as in MOD-SULQ is not a good approach. However, in lower dimensions and data-rich regimes, the performance may be more favorable.

A little calculation yields the sample complexity bound.

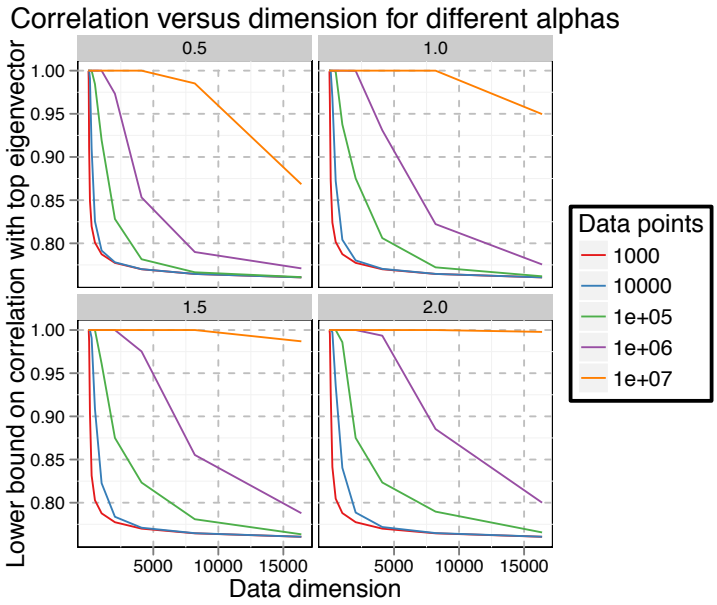

Figure 3: Upper bound on the correlation between $\langle \hat{v}_1, v_1 \rangle$ for MOD-SULQ. The horizontal axis is the dimension d of the data, and the vertical axis is the correlation. The four panels correspond to values of $\alpha = 0.5$, 1, 1.5, and 2.

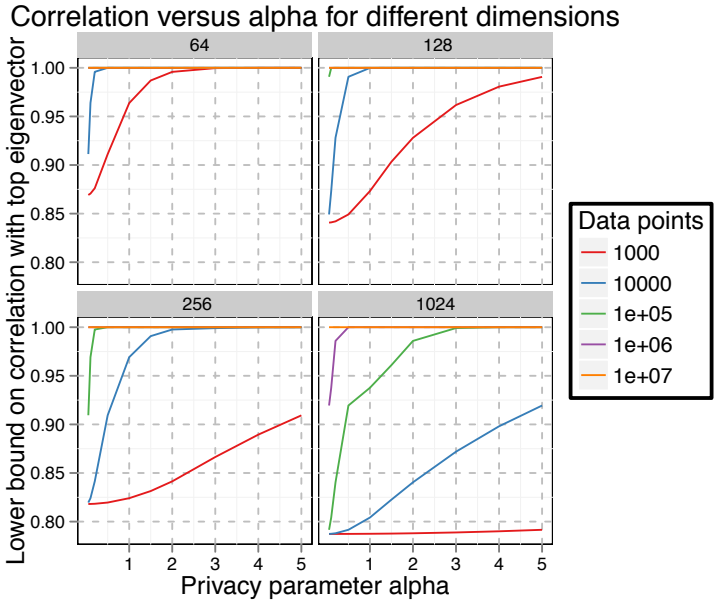

Figure 4: Upper bound on the correlation between $\langle \hat{v}_1, v_1 \rangle$ for MOD-SULQ. The horizontal axis is the privacy parameter α, and the vertical axis is the correlation. The four panels correspond to values of $d = 64$, 128, 256, and 1024.

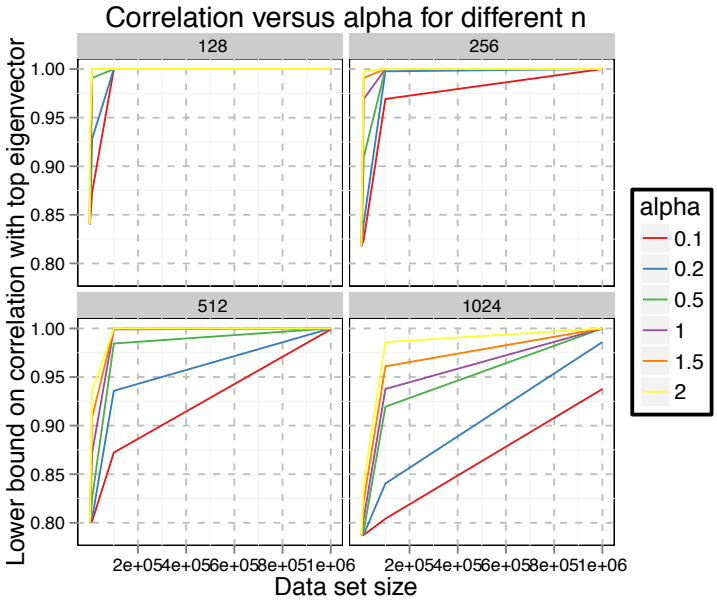

Figure 5: Upper bound on the correlation between $\langle \hat{v}_1, v_1 \rangle$ for MOD-SULQ. The horizontal axis is the size of the data set n, and the vertical axis is the correlation. The four panels correspond to values of $d = 64$, 128, 256, and 1024.

Proof of Theorem 5. Suppose $\mathbb{E}\left[|\langle \hat{v}_1, v_1 \rangle|\right] = \rho$. Then a little algebra shows

$$2\sqrt{1-\rho} \geq \min_{\phi \in \Phi} \sqrt{1-\phi} \left(1 - \frac{1/\beta^2 + \log 2}{(d-1)\log \frac{1}{\sqrt{1-\phi^2}} - \log(8)} \right).$$

Set ϕ such that $(d-1)\log \frac{1}{\sqrt{1-\phi^2}} - \log(8) = 2(1/\beta^2 + \log 2)$ to get

$$4\sqrt{1-\rho} \geq \sqrt{1-\phi}.$$

Since we are concerned with the scaling behavior for large d and n,

$$\log \frac{1}{\sqrt{1-\phi^2}} = \Theta\left(\frac{1}{\beta^2 d}\right),$$

so

$$\begin{aligned} \phi &= \sqrt{1 - \exp\left(-\Theta\left(\frac{1}{\beta^2 d}\right)\right)} \\ &= \Theta\left(\sqrt{\frac{1}{\beta^2 d}}\right). \end{aligned}$$

Lower bound β in Algorithm 1 to get for some constant c_1,

$$\beta^2 > c_1 \frac{d^2}{n^2 \alpha^2} \log(d/\delta).$$

Substituting this we get for some constant c_2 that

$$(1 - c_2(1-\rho)) \leq c_3 \frac{n^2\alpha^2}{d^3 \log(d/\delta)}.$$

Now solving for n shows

$$n \geq c \frac{d^{3/2}\sqrt{\log(d/\delta)}}{\alpha} \left(1 - c'(1-\rho)\right).$$

□

D Proof of Theorem 4

We now turn to a general lower bound on the sample complexity for any differentially private approximation to PCA. We construct K databases which differ in a small number of points whose top eigenvectors are not too far from each other. For such a collection, Lemma 7 shows that for any differentially private mechanism, the average correlation over the collection cannot be too large. That is, any α-differentially private mechanism cannot have high utility on all K data sets. The remainder of the argument is to construct these K data sets.

The proof uses some simple eigenvalue and eigenvector computations. A matrix of positive entries

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \tag{42}$$

has characteristic polynomial

$$\det(A - \lambda I) = \lambda^2 - (a+c)\lambda + (ac - b^2)$$

and eigenvalues

$$\begin{aligned} \lambda &= \frac{1}{2}(a+c) \pm \frac{1}{2}\sqrt{(a+c)^2 - 4(ac-b^2)} \\ &= \frac{1}{2}(a+c) \pm \frac{1}{2}\sqrt{(a-c)^2 + 4b^2}. \end{aligned} \tag{43}$$

The eigenvectors are in the directions $(b, -(a-\lambda))^T$.

Lemma 7. *Let $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$ be K databases which differ in the value of at most $\frac{\ln(K-1)}{\alpha}$ points, and let $u_1, \ldots, u_K$ be the top eigenvectors of $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K$. If $\mathcal{A}$ is any α-differentially private algorithm, then,*

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[|\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] \leq K \left(1 - \frac{1}{16}(1 - \max |\langle u_i, u_j \rangle|) \right).$$

Proof. Let

$$t = \min_{i \neq j} (\|u_i - u_j\|, \|u_i + u_j\|),$$

and $\mathcal{G}_i$ be the cap around $\pm u_i$ of radius $t/2$:

$$\mathcal{G}_i = \{u : \|u - u_i\| < t/2\} \cup \{u : \|u + u_i\| < t/2\}.$$

We claim that

$$\sum_{i=1}^{K} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \notin G_i) \geq \frac{1}{2}(K-1). \tag{44}$$

The proof is by contradiction. Suppose the claim is false. Because all of the caps $\mathcal{G}_i$ are disjoint, and applying the definition of differential privacy,

$$\begin{aligned} \frac{1}{2}(K-1) &> \sum_{i=1}^{K} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \notin G_i) \\ &\geq \sum_{i=1}^{K} \sum_{i' \neq i} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \in G_{i'}) \\ &\geq \sum_{i=1}^{K} \sum_{i' \neq i} e^{-\alpha \cdot \ln(K-1)/\alpha} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_{i'}) \in G_{i'}) \\ &\geq (K-1) \cdot \frac{1}{K-1} \cdot \sum_{i=1}^{K} \mathbb{P}_{\mathcal{A}}(\mathcal{A}(\mathcal{D}_i) \in G_i) \\ &\geq K - \frac{1}{2}(K-1), \end{aligned}$$

which is a contradiction, so (44) holds. Therefore by the Markov inequality

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[\min(\|\mathcal{A}(\mathcal{D}_i) - u_i\|^2, \|\mathcal{A}(\mathcal{D}_i) + u_i\|^2) \right] \geq \sum_{i=1}^{K} \mathbb{P}(\mathcal{A}(\mathcal{D}_i) \notin G_i) \cdot \frac{t^2}{4}$$
$$\geq \frac{1}{8}(K-1)t^2.$$

Rewriting the norms in terms of inner products shows

$$2K - 2\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[|\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] \geq \frac{1}{8}(K-1)\left(2 - 2\max |\langle u_i, u_j \rangle|\right),$$

so

$$\sum_{i=1}^{K} \mathbb{E}_{\mathcal{A}} \left[|\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] \leq K \left(1 - \frac{1}{8}\frac{K-1}{K}(1 - \max |\langle u_i, u_j \rangle|) \right)$$
$$\leq K \left(1 - \frac{1}{16}(1 - \max |\langle u_i, u_j \rangle|) \right).$$

□

Proof of Theorem 4. From Lemma 7, given a set of K databases differing in $\frac{\ln(K-1)}{\alpha}$ points with top eigenvectors $\{u_i : i = 1, 2, \ldots, K\}$, for at least one database i,

$$\mathbb{E}_{\mathcal{A}} \left[|\langle \mathcal{A}(\mathcal{D}_i), u_i \rangle| \right] \leq 1 - \frac{1}{16}\left(1 - \max |\langle u_i, u_j \rangle|\right)$$

for any α-differentially private algorithm. Setting the left side equal to some target ρ,

$$1 - \rho \geq \frac{1}{16}\left(1 - \max |\langle u_i, u_j \rangle|\right). \tag{45}$$

So our goal is construct these data bases such that the inner product between their eigenvectors is small.

Let $y = e_d$, the d-th coordinate vector, and let $\phi \in ((2\pi d)^{-1/2}, 1)$. Lemma 5 shows that there exists a packing $\mathcal{W} = \{w_1, w_2, \ldots, w_K\}$ of the sphere $\mathbb{S}^{d-2}$ spanned by $\{e_1, e_2, \ldots, e_{d-1}\}$ such that $\max_{i \neq j} |\langle w_i, w_j \rangle| \leq \phi$, where

$$K = \frac{1}{8}(1 - \phi)^{-(d-2)/2}.$$

Choose ϕ such that $\ln(K-1) = d$. This means

$$1 - \phi = \exp\left(-2 \cdot \frac{\ln 8 + \ln(1 + \exp(d))}{d-2} \right).$$

The right side is minimized for $d = 3$ but this leads to a rather weak lower bound $1 - \phi > 3.5 \times 10^{-5}$. By contrast, for $d = 100$, the bound is $1 - \phi > 0.12$. In all cases, $1 - \phi$ is at least a constant value.

We will construct one database for each w_i. Let $\beta = \frac{d}{n\alpha}$. For now, we assume that $\beta \leq \Delta \leq \frac{1}{2}$. The other case, when $\beta \geq \Delta$ will be considered later. Because $\beta \leq \Delta$, we have

$$n > \frac{d}{\Delta\alpha}.$$

Each database will contain n points and they will differ in $n\beta = \frac{\ln(K-1)}{\alpha}$ points. The construction uses a parameter $0 \leq m \leq 1$ that will be set as a function of the eigenvalue gap Δ. We will derive conditions on n based on the requirements on d, α, ρ, and Δ. For $i = 1, 2, \ldots, K$ let the data set $\mathcal{D}_i$ contain

- $n(1 - \beta)$ copies of $\sqrt{m}y$

- $n\beta$ copies of $z_i = \frac{1}{\sqrt{2}}y + \frac{1}{\sqrt{2}}w_i$.

Thus datasets $\mathcal{D}_i$ and $\mathcal{D}_j$ differ in the values of $n\beta = \frac{\ln(K-1)}{n\alpha}$ individuals. The second moment matrix A_i of $\mathcal{D}_i$ is

$$A_i = ((1-\beta)m + \frac{1}{2}\beta)yy^T + \frac{1}{2}\beta(w_i^T y + yw_i^T) + \frac{1}{2}\beta w_i w_i^T.$$

By choosing an basis containing y and w_i, we can write this as

$$A_i = \begin{bmatrix} (1-\beta)m + \frac{1}{2}\beta & \frac{1}{2}\beta & \mathbf{0} \\ \frac{1}{2}\beta & \frac{1}{2}\beta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

This is in the form (42), with $a = (1-\beta)m + \frac{1}{2}\beta$, $b = \frac{1}{2}\beta$, and $c = \frac{1}{2}\beta$.

The matrix A_i has two nonzero eigenvalues given by

$$\lambda = \frac{1}{2}(a+c) + \frac{1}{2}\sqrt{(a-c)^2 + 4b^2}, \tag{46}$$

$$\lambda' = \frac{1}{2}(a+c) - \frac{1}{2}\sqrt{(a-c)^2 + 4b^2}, \tag{47}$$

The gap Δ between the top two eigenvalues is:

$$\Delta = \sqrt{(a-c)^2 + 4b^2} = \sqrt{m^2(1-\beta)^2 + \beta^2}.$$

We can thus set m in the construction to ensure an eigengap of Δ:

$$m = \frac{\sqrt{(\Delta^2 - \beta^2)}}{1-\beta}. \tag{48}$$

The top eigenvector of A_i is given by

$$u_i = \frac{b}{\sqrt{b^2 + (a-\lambda)^2}}y + \frac{(a-\lambda)}{\sqrt{b^2 + (a-\lambda)^2}}w_i.$$

where λ is given by (46). Therefore

$$\begin{aligned} \max_{i \neq j} |\langle u_i, u_j \rangle| &\leq \frac{b^2}{b^2 + (a-\lambda)^2} + \frac{(a-\lambda)^2}{b^2 + (a-\lambda)^2} \max_{i \neq j} |\langle w_i, w_j \rangle| \\ &\leq 1 - \frac{(a-\lambda)^2}{b^2 + (a-\lambda)^2}(1-\phi). \end{aligned} \tag{49}$$

To obtain an upper bound on $\max_{i \neq j} |\langle u_i, u_j \rangle|$ we must lower bound $\frac{(a-\lambda)^2}{b^2 + (a-\lambda)^2}$.

Since $x/(\nu + x)$ is monotonically increasing in x when $\nu > 0$, we will find a lower bound on $(a - \lambda)$. Observe that from (46),

$$\lambda - a = \frac{b^2}{\lambda - c}.$$

So to lower bound $\lambda - a$ we need to upper bound $\lambda - c$. We have

$$\lambda - c = \frac{1}{2}(a-c) + \frac{1}{2}\Delta = \frac{1}{2}\left((1-\beta)m + \Delta\right)$$

Because $b = \beta/2$,

$$(\lambda - a)^2 > \left(\frac{\beta^2}{2((1-\beta)m + \Delta)}\right)^2 = \frac{\beta^4}{4((1-\beta)m + \Delta)^2}.$$

Now,

$$\begin{aligned}\frac{(a-\lambda)^2}{b^2+(a-\lambda)^2} &> \frac{\beta^4}{\beta^2((1-\beta)m+\Delta)^2+\beta^4} \\ &= \frac{\beta^2}{\beta^2+((1-\beta)m+\Delta)^2} \\ &> \frac{\beta^2}{5\Delta^2}, \end{aligned} \tag{50}$$

where the last step follows by plugging in m from (48) and using the fact that $\beta \leq \Delta$. Putting it all together, we have from (45), (49), and (50), and using the fact that ϕ is such that $\ln(K-1) = d$ so that $\beta = \frac{d}{n\alpha}$,

$$\begin{aligned}1-\rho &\geq \frac{1}{16} \cdot \frac{(a-\lambda)^2}{b^2+(a-\lambda)^2}(1-\phi) \\ &> \frac{1-\phi}{80}\frac{\beta^2}{\Delta^2} \\ &= \frac{1-\phi}{80} \cdot \frac{d^2}{\Delta^2 n^2 \alpha^2}, \end{aligned}$$

which implies

$$n > \frac{\sqrt{1-\phi}}{\sqrt{80}} \cdot \frac{d}{\Delta\alpha\sqrt{1-\rho}}.$$

Thus for $\beta \leq \Delta \leq 1/2$, any α-differentially private algorithm needs $\Omega\left(\frac{d}{\Delta\alpha\sqrt{1-\rho}}\right)$ points to get expected inner product ρ on all data sets with eigengap Δ.

We now consider the case where $\beta > \Delta$. We choose a slightly different construction here. The i-th database now consists of $n(1-\beta)$ copies of the 0 vector, and $n\beta$ copies of $\frac{\Delta}{\beta}w_i$. Thus, every pair of databases differ in the values of $n\beta = \frac{\ln(K-1)}{\alpha}$ people, and the eigenvalue gap between the top two eigenvectors is $\beta \cdot \frac{\Delta}{\beta} = \Delta$.

As the top eigenvector of the i-th database is $u_i = w_i$,

$$\max_{i \neq j} |\langle u_i, u_j \rangle| = \max_{i \neq j} |\langle w_i, w_j \rangle| \leq \phi$$

Combining this with (45), we obtain

$$1-\rho \geq \frac{1}{16}(1-\phi),$$

which provides the additional condition in the Theorem. □

E Experiments and implementation

E.1 Description of the data sets

The `kddcup99` [17] contains features about 494,021 network connections, `census` [2] is a demographic data set on 199,523 individuals, `localization` [21] is a medical dataset with 164,860 instances of sensor readings on individuals engaged in different activities, and `insurance` [36] is a dataset on product usage and demographics of 9,822 individuals. We chose values of k such that the top-k PCA subspace had $q_{\mathrm{F}}(V)$ at least 80% of $\|A\|_{\mathrm{F}}$. A summary is in Table E.1.

All datasets contain a mix of continuous and categorical features. We preprocess each dataset by converting a feature with q discrete values to a vector in $\{0,1\}^q$; after preprocessing, the datasets `kddcup99`, `census`, `localization` and `insurance` have dimensions 116, 513, 44 and 150 respectively. We also normalize each column so that each entry has maximum value 1, and normalize each row such that the maximum (Euclidean) row norm is 1. We choose $k = 4$ for `kddcup`, $k = 8$ for `census`, $k = 10$ for `localization` and $k = 11$ for `insurance`; in each case, the utility $q_{\mathrm{F}}(U_k)$ of the top-k PCA subspace of the data matrix accounts for at least 80% of $\|A\|_{\mathrm{F}}$. Thus, all four datasets, although fairly high dimensional, have good low-dimensional representations. The properties of each dataset are summarized in Table E.1.

Dataset	#instances	#dimensions	k	$q_{\mathrm{F}}(U)/\left\|A\right\|_{\mathrm{F}}$
kddcup	494,021	116	4	0.96
census	199,523	513	8	0.81
localization	164,860	44	10	0.81
insurance	9,822	150	11	0.81

Table 2: Parameters of each dataset. The second column is the number of dimensions after preprocessing. k is the dimensionality of the PCA, and the fourth column contains $q_{\mathrm{F}}(U)/\left\|A\right\|_{\mathrm{F}}$ where U is the top k PCA subspace.

E.2 Implementation for classification

We used a linear SVM for all classification tasks, which is implemented by libSVM [6].

E.3 Implementation of Gibbs sampling

There is a mismatch between the theoretical analysis of differentially private algorithms and their implementation in real systems. Because differential privacy is a mathematical definition, the description of differentially private procedure makes a number of idealizations regarding computation. Some of these idealizations are related to running the algorithm in a real environment; the designers of differentially private systems such as Airavat [33] require additional security assumptions that have to be verified. At a more basic level, the difference between truly random noise and pseudorandomness [29, 27] and the effects of finite precision can lead to a gap between the theoretical ideal and practice. Finally, implementation of private algorithms can lead to further gaps between theory and practice. For example, implementing objective perturbation [8] uses numerical optimization tools for approximate solutions to convex optimization problems, which have complex termination conditions that are not part of the accompanying theoretical analysis. In this work, MCMC sampling does not sample exactly from the Bingham distribution, and we leave a theoretical investigation of the impact of approximate sampling for future work.[2]

The theoretical analysis of PPCA uses properties of the Bingham distribution $\mathsf{BMF}_k(\cdot)$ given in (8). To implement this algorithm for experiments we use a Gibbs sampler due to Hoff [19]. The Gibbs sampling scheme induces a Markov Chain, the stationary distribution of which is the density in (8). Gibbs sampling and other MCMC procedures are widely used in statistics, scientific modeling, and machine learning to estimate properties of complex distributions.

Finding the speed of convergence of MCMC methods is still an open area of research. There has been much theoretical work on estimating convegence times [18, 13, 20, 30, 31, 32, 22, 23], but unfortunately, most theoretical guarantees are available only in special cases and are often too weak for practical use. In lieu of theoretical guarantees, users of MCMC methods empirically estimate the *burn-in time*, or the number of iterations after which the chain is sufficiently close to its stationary distribution. Statisticians employ a range of diagnostic methods and statistical tests to empirically determine if the Markov chain is close to stationarity [10, 5, 4, 14]. These tests do not provide a sufficient guarantee of stationarity, and there is no "best test" to use. In practice, the convergence of derived statistics is used to estimate an appropriate the burn-in time. In the case of the Bingham distribution, Hoff [19] performs qualitative measures of convergence. Developing a better characterization of the convergence of this Gibbs sampler is also an important question for future work.

To choose an appropriate burn-in time, we examined the *time series trace* of the Markov Chain. We ran l copies of the chain, starting from l different initial locations drawn uniformly from the set of all $d \times k$ matrices with orthonormal columns. Let $X^i(t)$ be the output of the i-th copy at iteration t, and let U be the top k PCA subspace of the data. For each i, we plot the magnitude of the projection of $X^i(t)$ onto U. After a number of iterations, the projections should converge to the same value.

For each copy, we also plot the following statistic as a function of iteration T:

$$F_k^i(T) = \frac{1}{\sqrt{k}} \left\| \frac{1}{T} \sum_{t=1}^{T} X^i(t) \right\|_{\mathrm{F}},$$

[2]This paragraph also appears in the main text of the document.

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. The matrix Bingham distribution has mean 0, and hence with increasing T, the statistic $F_k^i(T)$ should converge to 0.

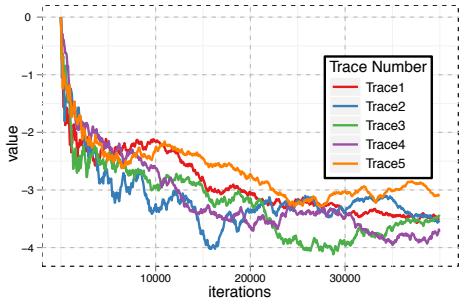

Figure 6: Plot of $\log F_k(T)$ for $k = 4$ as a function of iteration T for 40,000 iterations of the Gibbs sampler for the `kddcup` dataset.

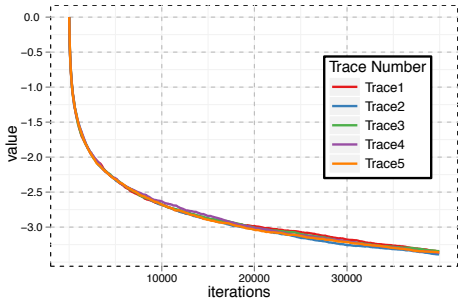

Figure 7: Plot of $\log F_k(T)$ for $k = 11$ as a function of iteration T for 40,000 iterations of the Gibbs sampler for the `insurance` dataset.

In our simulations, we observed that the Gibbs sampler converges to $F_k(t) < 0.01$ at $t = 20{,}000$ when run on data with a few hundred dimensions and with k between 5 and 10; we thus chose to run the Gibbs sampler for $T = 20{,}000$ timesteps for all the datasets. We show $\log F_k^i(T)$ as a function of iteration T for datasets `insurance` and `kddcup` in Figures 7 and 6 respectively; the plots are over 5 trajectories of the Markov Chain, which are initialized at 5 locations drawn uniformly from the set of all $d \times k$ matrices with orthonormal columns. The plots show that $F_k^i(T)$ decreases rapidly after a few thousand iterations, and is less than 0.01 after $T = 20{,}000$ in both cases. $\log F_k^i(T)$ also appears to have a larger variance for `kddcup` than for `insurance`; this is explained by the fact that the `kddcup` dataset has a much larger number of samples, which makes its stationary distribution farther from the initial distribution of the sampler.

Our simulations indicate that the chains converge fairly rapidly, particularly when $\|A - A_k\|_{\mathrm{F}}$ is small so that A_k is a good approximation to A. Convergence is slower for larger n when the initial state is chosen from the uniform distribution over all $k \times d$ matrices with orthonormal columns; this is explained by the fact that for larger n, the stationary distribution is farther in variation distance from the starting distribution, which results in a longer convergence time.

References

[1] AGRAWAL, R., AND SRIKANT, R. Privacy-preserving data mining. *SIGMOD Rec. 29*, 2 (2000), 439–450.

[2] ASUNCION, A., AND NEWMAN, D. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2007.

[3] BALL, K. An elementary introduction to modern convex geometry. In *Flavors of Geometry*, S. Levy, Ed., vol. 31 of *Mathematical Sciences Research Institute Publications*. Cambridge University Press, 1997, pp. 1–58.

[4] BROOKS, S. P., AND GELMAN, A. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics 7*, 4 (Dec. 1998), 434.

[5] BROOKS, S. P., AND ROBERTS, G. O. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing* (1998).

[6] CHANG, C.-C., AND LIN, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[7] CHAUDHURI, K., AND MISHRA, N. When random sampling preserves privacy. In *CRYPTO* (2006), C. Dwork, Ed., vol. 4117 of *Lecture Notes in Computer Science*, Springer, pp. 198–213.

[8] CHAUDHURI, K., MONTELEONI, C., AND SARWATE, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research 12* (March 2011), 1069–1109.

[9] CONWAY, J., AND SLOANE, N. *Sphere Packing, Lattices and Groups*. Springer-Verlag, New York, 1998.

[10] COWLES, M. K., AND CARLIN, B. P. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association 91*, 434 (June 1996), 883.

[11] CSISZÁR, I., AND NARAYAN, P. The capacity of the arbitrarily varying channel revisited : Positivity, constraints. *IEEE Transactions on Information Theory 34*, 2 (1988), 181–193.

[12] CSISZÁR, I., AND NARAYAN, P. Capacity of the Gaussian arbitrarily varying channel. *IEEE Transactions on Information Theory 37*, 1 (1991), 18–26.

[13] DOUC, R., MOULINES, E., AND ROSENTHAL, J. S. Quantitative bounds on convergence of time-inhomogeneous Markov chains. *The Annals of Applied Probability 14*, 4 (Nov. 2004), 1643–1665.

[14] ELADLOUNI, S., FAVRE, A., AND BOBEE, B. Comparison of methodologies to assess the convergence of Markov chain Monte Carlo methods. *Computational Statistics & Data Analysis 50*, 10 (June 2006), 2685–2701.

[15] EVFIMIEVSKI, A., GEHRKE, J., AND SRIKANT, R. Limiting privacy breaches in privacy preserving data mining. In *PODS* (2003), pp. 211–222.

[16] HAY, M., LI, C., MIKLAU, G., AND JENSEN, D. Accurate estimation of the degree distribution of private networks. In *ICDM* (2009), pp. 169–178.

[17] HETTICH, S., AND BAY, S. The UCI KDD Archive. University of California, Irvine, Department of Information and Computer Science, 1999.

[18] HOBERT, J. P., AND JONES, G. L. Sufficient burn-in for Gibbs samplers for a hierarchical random effects model. *The Annals of Statistics 32*, 2 (Apr. 2004), 784–817.

[19] HOFF, P. D. Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data. *J. Comp. Graph. Stat. 18*, 2 (2009), 438–456.

[20] JONES, G. L., AND HOBART, J. P. Honest Exploration of Intractable Probability Distributions via Markov Chain Monte Carlo. *Statistical Science 16*, 4 (2001), 312–334.

[21] KALUZA, B., MIRCHEVSKA, V., DOVGAN, E., LUSTREK, M., AND GAMS, M. An agent-based approach to care in independent living. In *International Joint Conference on Ambient Intelligence (AmI-10)* (2010).

[22] KOLASA, J. E. Convergence and Accuracy of Gibbs Sampling for Condirtional Distributions in Generalized Linear Models. *The Annals of Statistics 27*, 1 (1999), 129–142.

[23] KOLASA, J. E. Explicit Bounds for Geometric Covergence of Markov Chains. *Journal of Applied Probability 37*, 3 (2000), 642–651.

[24] LI, N., LI, T., AND VENKATASUBRAMANIAN, S. Closeness: A new privacy measure for data publishing. *IEEE Trans. Knowl. Data Eng. 22*, 7 (2010), 943–956.

[25] MACHANAVAJJHALA, A., GEHRKE, J., KIFER, D., AND VENKITASUBRAMANIAM, M. l-diversity: Privacy beyond k-anonymity. In *ICDE* (2006), p. 24.

[26] MACHANAVAJJHALA, A., KIFER, D., ABOWD, J. M., GEHRKE, J., AND VILHUBER, L. Privacy: Theory meets practice on the map. In *ICDE* (2008), pp. 277–286.

[27] MCGREGOR, A., MIRONOV, I., PITASSI, T., REINGOLD, O., TALWAR, K., AND VADHAN, S. The limits of two-party differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS '10)* (October 2010), pp. 23–26.

[28] MCSHERRY, F., AND TALWAR, K. Mechanism design via differential privacy. In *FOCS* (2007), pp. 94–103.

[29] MIRONOV, I., PANDEY, O., REINGOLD, O., AND VADHAN, S. Computational differential privacy. In *Advances in Cryptology - CRYPTO 2009*, S. Halevi, Ed., vol. 5677 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2009, pp. 126–142.

[30] ROBERTS, G. Bounds on regeneration times and convergence rates for Markov chains. *Stochastic Processes and their Applications 80*, 2 (Apr. 1999), 211–229.

[31] ROBERTS, G. O., AND SAHU, S. K. Approximate Predetermined Convergence Properties of the Gibbs Sampler. *Journal of Computational and Graphical Statistics 10*, 2 (June 2001), 216–229.

[32] ROSENTHAL, J. S. Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *Journal of the American Statistical Association 90*, 430 (June 1995), 558–566.

[33] ROY, I., SETTY, S. T. V., KILZER, A., SHMATIKOV, V., AND WITCHEL., E. Airavat: Security and privacy for mapreduce. In *Proceedings of the 7th Usenix Symposium on Networked Systems Design and Implementation (NSDI)* (2010).

[34] SHANNON, C. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal 38* (1959), 611–656.

[35] SWEENEY, L. k-anonymity: a model for protecting privacy. *Int. J. on Uncertainty, Fuzziness and Knowledge-Based Systems* (2002).

[36] VAN DER PUTTEN, P., AND VAN SOMEREN, M. Coil challenge 2000: The insurance company case, 2000. Leiden Institute of Advanced Computer Science Technical Report 2000-09.

[37] YU, B. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. L. Yang, Eds., Research Papers in Probability and Statistics. Springer-Verlag, 1997, ch. 29, pp. 423–425.