CS229r: Mathematical Approaches to Data Privacy
Homework 3

Due Friday, April 5th, 2013 by 5PM.
(Email to `diffprivcourse-hw@seas.harvard.edu`)

- You must type your solutions. LaTeX or Microsoft Word are both fine. Let us know if there is something else you prefer to use. If you use LaTeX, please submit both the compiled PDF file and the source.
- Strive for clarity and conciseness in your solutions, emphasizing the main ideas over low-level details.
- Remember to cite all collaborators and sources of ideas.

Problem 1 (More on Local Sensitivity)**.** Recall that in the propose-test-release framework (Dwork-Roth Chapter 7.2), the approach is to propose a bound b on the local sensitivity of a function f with respect to the database x, test that the input database x is far from having local sensitivity larger than b in a differentially private way, and then release the noisy answer $f(x) + \mathrm{Lap}(b/\varepsilon)$ if the test succeeds or output $\perp$ if the test fails. Thus the algorithm either answers with error $O(b/\varepsilon)$ or outputs $\perp$. In this problem you will show this sort of error guarantee is essentially the best possible, in that any differentially private algorithm can only answer the query to within error $\pm b$ on databases such that the local sensitivity is $O(b)$ on all nearby databases.

Suppose that we have an (ε, δ)-dp mechanism $M : \mathcal{X}^n \to \mathbb{R}$ such that for every x,

$$\Pr\left[M(x) \in [f(x) - b, f(x) + b] \cup \{\perp\}\right] \geq 1 - o(1).$$

1. Show that if the local sensitivity of f with respect to x is strictly larger than $2b$, $\varepsilon = O(1)$, and $\delta = o(1)$, then
$$\Pr\left[M(x) = \perp\right] \geq 1 - o(1).$$
2. Show that if there is a database x' such that $|x \triangle x'| \leq 1/\varepsilon$ and the local sensitivity of f with respect to x' is strictly larger than $4b$, $\varepsilon = O(1)$, and $\delta = o(\varepsilon)$. then
$$\Pr\left[M(x) = \perp\right] \geq 1 - o(1).$$

Problem 2 (More on Generating Synthetic Data)**.**

1. In this question, you'll show that we can get efficient synthetic data for conjunctions if we assume the database comes from a "nice" distribution. Consider the following family of distributions over databases, $\mathcal{D}_{\mathbf{p}}$: for each $j = 1, \ldots, d$, there is some *bias* p_j. For each row

$x_i \in \{0,1\}^d$, each bit x_{ij} is randomly and independently generated to be 1 with probability p_j and 0 otherwise. The database x then consists of these n rows.

Show that there is an ε-dp mechanism $M : (\{0,1\}^d)^n \to (\{0,1\}^d)^{\widehat{n}}$ (for $n \geq (\text{poly}(d)/\varepsilon$ and $\widehat{n}$ of your choosing) that runs in time $\text{poly}(n,d)$ and generates accurate synthetic data for all conjunction queries with high probability. That is, give a mechanism M such that for every $\mathbf{p} = (p_1, \ldots, p_d)$, if the database is drawn $x \leftarrow_{\text{R}} \mathcal{D}_{\mathbf{p}}$ and we draw $\widehat{x} \leftarrow_{\text{R}} M(x)$, then with probability at least .99,

$$\forall q \in Q, |q(x) - q(\widehat{x})| \leq .01$$

where Q is the set of all conjunction queries. Note that the differential privacy should hold for *every* database, regardless of whether or not it was sampled according to $\mathcal{D}_{\mathbf{p}}$.

2. Say that a mechanism $M : (\{0,1\}^d)^n \times 2^Q \to (\{0,1\})^d)^{\widehat{n}}$ generates *α-weakly-accurate synthetic data wrt to Q* if for every $x \in (\{0,1\}^d)^n$ and $Q' \subseteq Q$, $M(x, Q')$ outputs $\widehat{x}$ such that with probability at least .99,

$$\frac{1}{|Q'|} \sum_{q \in Q'} |q(x) - q(\widehat{x})| \leq \alpha.$$

Recall that the definition in the reading and in class required that (with high probability) the synthetic database $\widehat{x}$ answered *every* query up to error α, whereas here we only require that the average error be at most.

Show that the hardness results for synthetic data still apply to mechanisms that generate weakly accurate private synthetic data. Specifically, show that for some sufficiently small constant $\alpha > 0$, there is no differentially private mechanism M that runs in time $\text{poly}(n,d)$ and generates α-weakly accurate synthetic data wrt $Q = \{\text{2-literal conjunctions}\}$.

Problem 3 (Faster Algorithms for Releasing Marginals on Sparse Databases)**.** Many databases that arise in practice are *sparse*, roughly meaning that most of the entries in the database are 0. In this problem you will show how to answer all conjunction queries on sparse databases more efficiently and using less data than we know how to achieve for arbitrary databases. Specifically, we will call a database $(\{0,1\}^d)^n$ *s-sparse* if

$$\frac{1}{n} \sum_{i=1}^{n} |x_i| \leq s.$$

Here $|x_i|$ is the Hamming weight of x_i, that is, the number of 1's in x_i.

Show that there is an ε-differentially private mechanism M with running time $n \cdot d^{O(\sqrt{s})}$ that on input an s-sparse database $x \in (\{0,1\}^d)^n$, releases a summary that enables computing any monotone conjunction query on x up to an additive error at most ± 0.01 as long as $n \geq d^{O(\sqrt{s})}/\varepsilon$. Note that your mechanism should be ε-dp for *every* input database x.