Question 1

In this question you will show that the database reconstruction algorithm from Lecture 6 can be made efficient.

We will say that a vector $y \in [-2,2]^m$ is *β-heavy* if at least $m/10$ of its coordinates have absolute value at least β. Let

$$q'_S(y) = \sum_{i \in S} y_i - \sum_{i \notin S} y_i$$

where S is a subset of $[m]$ and y is a vector in $\mathbb{R}^m$.

(a) Show that if $y \in [-2,2]^m$ is 1/4-heavy and S is a random subset of $[m]$, then there exists a sufficiently small constant γ (independent of m) such that

$$\Pr[q'_S(y) \geq \gamma\sqrt{m}] \geq \gamma.$$

Solution: We can write $q'_S(y) = X = \sum_{i=1}^m X_i y_i$ where $X_1, \dots, X_m$ are i.i.d. $\{-1,1\}$ random variables. Then $\mathrm{E}[X] = 0$, $\mathrm{E}[X^2] = \sum_{i=1}^m y_i^2 \geq (m/10)(1/16) \geq m/160$, and $\mathrm{E}[X^4] = \sum_{i=1}^m y_i^4 + \sum_{i \neq j} 3y_i^2 y_j^2 \leq 16m + 48m(m-1) \leq 48m^2$. By the Paley-Zygmund inequality,

$$\Pr[X \geq \sqrt{m}/60] \geq \Pr[X^2 \geq \tfrac{1}{4}\mathrm{E}[X^2]] \geq \frac{9}{16} \cdot \frac{(m/160)^2}{(48m^2)^2} \geq 10^{-9}.$$

(b) Let G be a finite subset of $[-1,1]^m$ and $\mathcal{S}$ be a collection of s random independent subsets of $[m]$. Show that the probability there exist $x \in \{-1,0,1\}^m$ and $x' \in G$ such that $x - x'$ is 1/4-heavy but $q'_S(x-x') < \gamma\sqrt{m}$ for all $S \in \mathcal{S}$ is at most $3^m|G|(1-\gamma)^s$.

Solution: For fixed x, x' such that $x - x'$ is 1/4 heavy and a single random subset S, by part (a) the probability that $q'_S(x-x') < \gamma\sqrt{m}$ is at most $1-\gamma$. By independence, the probability that there exists such an S in $\mathcal{S}$ is at most $(1-\gamma)^s$. Taking a union bound over at most 3^m choices of x and at most $|G|$ choices for x' gives the desired conclusion.

(c) Show that if $s \geq Km\log m$ for a sufficiently large constant K, then with probability at least 1/2 over the choice of $\mathcal{S}$, for every $x \in \{-1,0,1\}^m$ and every $x' \in [-1,1]^m$ such that $x - x'$ is 1/3-heavy, there exists a set $S \in \mathcal{S}$ such that $q'_S(x-x') \geq \gamma\sqrt{m}/2$. (**Hint:** Take G to be a sufficiently dense grid in $[-2,2]^m$.)

Solution: Let $D = \lceil \sqrt{m}/\gamma \rceil$ and let G be the set of all points of the form $(d_1/D, \dots, d_m/D)$ where $d_1, \dots, d_m$ are integers ranging from $-2D$ to $2D$. Then $|G| = (4D)^m = 2^{O(m \log m)}$. By part (b), for K sufficiently large, with probability at least 1/2 for every pair $x \in \{-1,0,1\}^m$ and $x^* \in G$ there exists a set $S \in \mathcal{S}$ such that $q'_S(x - x^*) \geq \gamma\sqrt{m}$. Assume this is the case and let $x, x' \in [-1,1]^m$ be any pair of points such that $x - x'$ is 1/3-heavy. If x^* is the closest point to x' in G (in ℓ_∞ distance) then $x - x^*$ must be 1/4 heavy because for any coordinate i,

$$|x_i - x_i^*| \geq |x_i - x_i'| - |x_i' - x_i^*| \geq |x_i - x_i'| - \frac{1}{12m}$$

so if $x_i - x_i' \geq 1/3$, $x_i - x_i^*$ must be at least 1/4. Then there exists a set S such that $q_S'(x - x^*) \geq \gamma\sqrt{m}$. For this set S,

$$q_S'(x - x') = q_S'(x - x^*) - q_S'(x^* - x') \geq \gamma\sqrt{m} - |q_S'(x^* - x')|.$$

The entries of $x^* - x'$ have value between $-1/2D$ and $1/2D$, so $|q_S'(x^* - x')| \leq m/2D \leq \gamma\sqrt{m}/2$, so $q_S'(x - x^*) \geq \gamma\sqrt{m}/2$ as desired.

(d) Suppose that M is a mechanism that on input[1] $x \in \{-1, 0, 1\}^m$ and query q_S' outputs an approximation to $q_S'(x)$ with additive error $\gamma\sqrt{m}/6$. Show that with constant probability, the following algorithm outputs a vector $\hat{x}$ that agrees with x on $9m/10$ of its coordinates:

(i) Choose a collection $\mathcal{S}$ of s independent uniform random subsets of $[m]$.

(ii) Query M to obtain approximations a_S to $q_S'(x)$ for all $S \in \mathcal{S}$.

(iii) Find $x' \in [-1, 1]^m$ such that $|q_S'(x') - a_S| \leq \gamma\sqrt{m}/6$, if it exists.
(This is a linear program; it can be solved efficiently.)

(iv) For every coordinate i, set

$$\hat{x}_i = \begin{cases} 1, & \text{if } x_i' \geq 1/2, \\ -1, & \text{if } x_i' \leq 1/2, \\ 0, & \text{otherwise} \end{cases}$$

and output $\hat{x}$.

Solution: By assumption, $x' = x$ is always a feasible solution in step (iii), so the algorithm always finds some x'. On the other hand, any x' that the algorithm outputs must satisfy

$$|q_S'(x' - x)| \leq |q_S'(x') - a_S| + |a_S - q_S'(x)| \leq \frac{\gamma\sqrt{m}}{6} + \frac{\gamma\sqrt{m}}{6} = \frac{\gamma\sqrt{m}}{3}$$

for all $S \in \mathcal{S}$. By part (c), $x - x'$ cannot be 1/3-heavy, so at least $9m/10$ coordinates of $x - x'$ have absolute value less than 1/3. On each of these coordinates, $\hat{x}_i$ must equal x, so $\hat{x}$ and x match on $9m/10$ of their coordinates.

Question 2

In this question you will that if a synthetic database mechanism is differentially private then its output is unlikely to contain rows from the original database. Let $M\colon D^n \to D^d$ be a synthetic database mechanism.

[1] In the actual database, we include the row $(i, 1)$ if $x_i = 1$, $(i, -1)$ if $x_i = -1$, and do not include a row that starts with i otherwise.

(a) Let $x \in D^n$ be a database whose rows are independent uniform samples from D and x' be a database obtained by resampling the ith row of x uniformly from D and independently of the other rows. Show that

$$\Pr_{M,x,x'}[M(x') \text{ contains the } i\text{-th row of } x] \leq d/|D|.$$

Solution: Conditioned on $M(x')$ the i-th row of x, which we call x_i, is a uniform random row in$D$. For every j, the probability that x_i equals the j-th row of $M(x')$ is $1/|D|$. By a union bound over all rows of $M(x')$ we obtain the bound of $d/|D|$.

(b) Use part (a) to show that if M is (ε, δ)-differentially private, then

$$\Pr_{M,x,x'}[M(x) \text{ contains at least one row of } x] \leq e^{\varepsilon} dn/|D| + \delta n.$$

Solution: By differential privacy, for every i,

$$\Pr[M(x) \text{ contains } x_i] \leq e^{\varepsilon} \Pr[M(x') \text{ contains } x_i] + \delta \leq e^{\varepsilon} d/|D| + \delta.$$

Taking a union bound over all i proves the claim.

(c) Now let $\mathcal{D}$ be an arbitrary distribution over D and assume the rows of x and x' are sampled as in part (a), but from $\mathcal{D}$ instead of the uniform distribution over D. Show that

$$\Pr_{M,x,x'}[M(x) \text{ contains at least one row of } x] \leq e^{\varepsilon} pdn + \delta n.$$

where $p = \max_r\{\Pr_{R\sim\mathcal{D}}[R = r]\}$. (You do not need to redo the proofs from parts (a) and (b), just explain the differences.)

Solution: In part (a), the probability that x_i equals the j-th row of x' is no longer $1/|D|$, but it is at most p. The rest of the proof is exactly the same with all instances of $1/|D|$ replaced by p.

(d) **(Extra credit)** Now suppose x is chosen from the following distribution: The i-th row of x equals $(i, 0)$ with probability $1/2$ and $(i, 1)$ with probability $1/2$, independently from the other rows. If the output of $M(x)$ contains 99% of the rows of x with probability at least 99%, can M be $(0.1, n^{-100})$-differentially private for sufficiently large n?

Question 3

Let P be a subset of $\{0, 1\}^n$. A *testing algorithm* for property P is a randomized algorithm M such that $\Pr[M(x) \text{ accepts}] \geq 2/3$ for every $x \in P$ and $\Pr[M(x') \text{ accepts}] \leq 1/3$ for every $x' \in \{0, 1\}^n$ that differs from all $x \in P$ in at least εn coordinates.

(a) Show that every P has a $O(1/\varepsilon n)$-differentially private testing algorithm.

Solution: Let M be the exponential mechanism with outcomes accept and reject and utilities

$$u(x, \text{accept}) = -\min_{x' \in P} |x - x'| \quad \text{and} \quad u(x, \text{reject}) = -\min_{x' \notin P} |x - x'|.$$

Then u is 1-sensitive, so the exponential mechanism with parameter $1/\varepsilon n$ is $1/\varepsilon n$-differentially private.

If $x \in P$, then $u(x, \text{accept}) > u(x, \text{reject})$ so $M(x)$ accepts with probability at least $1/2$. If x differs from all $x' \in P$ in at least εn coordinates, then $u(x, \text{accept}) < -\varepsilon n$ and $u(x, (reject)) = 0$, so

$$\Pr[M(x) \text{ accepts}] < \frac{e^{-1}}{e^{-1} + e^0} < 0.269.$$

This does not quite meet the requirements, where the probabilities should be 1/3 and 2/3. One way to achieve this is to change the utilities to, say,

$$u(x, \text{accept}) = \varepsilon n/2 - \min_{x' \in P} |x - x'| \quad \text{and} \quad u(x, \text{reject}) = \varepsilon n/2 - \min_{x' \notin P} |x - x'|.$$

and use a slightly larger privacy parameter, say $3/\varepsilon n$, and repeat the same analysis.

(b) A testing algorithm is *one-sided* if $\Pr[M(x) \text{ accepts}] = 1$ for every $x \in P$. Which P have a $(100, 0.1)$-differentially private one-sided testing algorithm?

Solution: If you set $\Pr[M(x) \text{ accepts}]$ to equal one for $x \in P$, 0.9 for x that differ from some $x' \in P$ in one coordinate, 0.8 for x that differ from some x' in P in two coordinates, and so on, and 0 for the remaining x, the resulting algorithm is one-sided and differentially private. This is not what I meant to ask.

What I had meant to ask is which P have a 100-differentially private algorithm. Then if $M(x)$ rejects with probability 0 for any x, it is forced to reject with probability 0 for all x, so $M(x)$ accepts all inputs. It follows that every string in $\{0,1\}^n$ must be within distance εn of some string in P. In coding theory terminology, P is then a covering code of radius εn.