**CS 860: Algorithms for Private Data Analysis** Fall 2020

# Lecture 11 — Packing Lower Bounds

*Prof. Gautam Kamath* *Scribe: Gautam Kamath*

Up to this point, we've seen many different algorithms for differential privacy. This naturally raises the question of whether this is the best we can do, or if it's possible to do better. Today, we're going to see that in fact some of the simple algorithms we've seen so far are optimal (or at least nearly optimal).

For the sake of presentation we'll be focusing on two main classes of queries, though these techniques are in fact more general As usual, we are given a dataset $X \in \mathcal{X}^n$.

1. **One-way marginal queries.** Our data domain is $\mathcal{X} = \{0,1\}^d$. We wish to answer the set of $d$ queries $f_j(X) = \frac{1}{n}\sum_{i=1}^n X_i^{(j)}$ for all $j \in [d]$, where $X_i^{(j)}$ is the $j$th coordinate of the $i$th point in the dataset.
2. **Histograms.** Our data domain is $\mathcal{X} = [k]$. We wish to answer the set of $k$ queries $f_j(X) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i = j\}$, for all $j \in [k]$.

For both families of queries, the error is measured in $\ell_\infty$-error: that is, the maximum of how far off our answer is for any individual query.

For the problem of one-way marginals, the Laplace mechanism with basic and advanced composition give (roughly) the following upper bounds:

- Under $\varepsilon$-pure DP, if $n = \tilde{\Omega}(d/\varepsilon\alpha)$,[1] then we have error $\leq \alpha$ for $f_j(X)$ for all $j \in [d]$ with probability $\geq 1/2$.
- Under $(\varepsilon, \delta)$-approximate DP, if $n = \tilde{\Omega}(\sqrt{d\log(1/\delta)}/\varepsilon\alpha)$, then we have error $\leq \alpha$ for $f_j(X)$ for all $j \in [d]$ with probability $\geq 1/2$.

Observe that there is a gap of $\sqrt{d}$ between the two necessary values of $n$. Today, we will see a nearly-matching lower bound for the pure DP case, showing that this gap is inherent – pure DP costs more than approximate DP.

On the other hand, for histograms we have the following bounds, using the Laplace histogram and the stability-based histogram mentioned in the last lecture.

- Under $\varepsilon$-pure DP, if $n = \Omega(\log k/\varepsilon\alpha)$, then we have error $\leq \alpha$ for $f_j(X)$ for all $j \in [k]$ with probability $\geq 1/2$.
- Under $(\varepsilon, \delta)$-approximate DP, if $n = \Omega(\log(1/\delta)/\varepsilon\alpha)$, then we have error $\leq \alpha$ for $f_j(X)$ for all $j \in [k]$ with probability $\geq 1/2$.

[1] Note that $\tilde{O}$ disregards logarithmic factors in the argument.

We will see that the former sample complexity is tight, and no algorithm can achieve the same guarantees with a smaller value of $n$.

The style of argument will use a "packing" approach. The term packing comes from the idea of trying to construct many different datasets, which each give different answers from each other on the set of queries. To use a metaphor, consider packing many (rigid) balls into a box: every ball will be far from each other (at least as far as the radius of each ball). The more datasets we can pack, the stronger our lower bounds will be.

The argument, at its core, is quite elegant. If an algorithm is supposed to be *accurate*, our construction implies that it will have to give different answers for every database. On the other hand, if an algorithm is supposed to be *private*, group privacy implies that the distribution of answers for every database must be similar. Putting these two constraints together, we get a lower bound on how much data is necessary to achieve both of these properties simultaneously.

We illustrate this with a very simple example: trying to output the mode of a dataset over a data domain $\mathcal{X} = [k]$. We want an $\varepsilon$-DP algorithm $M$ which outputs the mode correctly with probability at least $1/2$. Consider the following set of $k$ datasets: $D_1, \ldots, D_k$, where dataset $D_j$ consists of $n$ copies of point $j \in [k]$. Fix some specific $j$, and consider the distribution of $M(D_j)$. Since there are $k$ possible outcomes, the pigeon-hole principle says that there must be at least one outcome $\Pr[M(D_j) = \ell] \leq 1/k$.[2] On the other hand, accuracy tells us that $\Pr[M(D_\ell) = \ell] \geq 1/2$. Note that we can convert from $D_\ell$ to $D_j$ by changing $n$ datapoints (i.e., the whole dataset). We will use this in combination with group privacy – while this regime might seem unusual, it's useful for proving lower bounds. In particular, it gives the following:

$$\frac{1}{2} \leq \Pr[M(D_\ell) = \ell] \leq e^{n\varepsilon} \Pr[M(D_j) = \ell] \leq \frac{e^{n\varepsilon}}{k}.$$

Taking the logarithm of both sides gives $n \geq \log(k/2)/\varepsilon$: saying we need $\Omega(\log k/\varepsilon)$ datapoints in order to achieve both accuracy and privacy simultaneously.

This is the core argument specialized to a simple case, we will derive results which apply in somewhat more general settings.

**Theorem 1.** *Let $D_1, \ldots, D_m \in \mathcal{X}^n$ be a set of $m$ datasets, which are at Hamming distance at most $t$ from some fixed dataset $D \in \mathcal{X}^n$. Let $Y_1, \ldots, Y_m \in \mathcal{Y}$ be a set of $m$ disjoint subsets of the space $\mathcal{Y}$. If there is an $\varepsilon$-DP mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ such that $\Pr[M(D_\ell) \in Y_\ell] \geq p$ for every $\ell \in [m]$, then*

$$\frac{1}{m} \geq pe^{-t\varepsilon}.$$

*Proof.* For any $\ell$, the accuracy guarantee says $\Pr[M(D_\ell) \in Y_\ell] \geq p$. Group privacy says that $\Pr[M(D_\ell) \in Y_\ell] \leq \Pr[M(D) \in Y_\ell]e^{t\varepsilon}$. Rearranging implies that $\Pr[M(D) \in Y_\ell] \geq pe^{-t\varepsilon}$.

Since the $Y_\ell$'s are disjoint, we have that

$$1 \geq \Pr[M(D) \in \cup_{\ell \in [m]} Y_\ell] = \sum_{\ell \in [m]} \Pr[M(D) \in Y_\ell] \geq mpe^{-t\varepsilon}.$$

Rearranging this gives the desired conclusion. □

[2] If not, then the sum of the probabilities would exceed 1.

To prove a lower bound for one-way marginals, we will focus on a simple case: $p = 1/2$ and $t = n$ (which is the trivial bound, since all datasets are at distance at most $n$ apart). In this case, we have $m \leq 2e^{n\varepsilon}$, or $n \geq \log(m/2)/\varepsilon$ – the same statement we showed in the toy example above.

Let us see how to apply this theorem to prove the following lower bound for one way marginals. In fact, for educational purposes, we will consider the fixed case when $\alpha = 1/2$.

**Theorem 2.** *Any $\varepsilon$-DP algorithm $M : \{0,1\}^n \to [0,1]^d$ which simultaneously answers all one-way marginals to error $< 1/2$ with probability $\geq 1/2$ requires $n = \Omega(d/\varepsilon)$.*

*Proof.* We consider the following set of $m = 2^d$ databases: for each point $w$ in $\{0,1\}^d$, each database consists of $n$ copies of the single point $w$. Let $D_w$ be the string corresponding to point $w$, and $Y_w$ is the $\ell_\infty$-ball of radius $1/2$ surrounding $w$:

$$Y_w = \{x \in [0,1]^d : |x_j - w_j| < 1/2, \ \forall j \in [k]\}.$$

Observe that these sets are indeed disjoint. Furthermore, they exactly correspond with our desired accuracy guarantee: we need that $\Pr[M(D_w) \in Y_w] \geq 1/2$. Applying Theorem 1 with these $m = 2^d$ databases and sets, fixing $p = 1/2$, and using $t = n$, we get $n = \Omega(\log(2^d)/\varepsilon) = \Omega(d/\varepsilon)$, as desired. □

Again, this shows that we need $n = \Omega(d)$ datapoints to compute the marginals to non-trivial accuracy under pure DP, whereas $n = O(\sqrt{d})$ are sufficient under approximate DP. It is also possible to show that $n = \Omega(\sqrt{d})$ samples are necessary for this problem under approximate DP, but this uses a significantly more challenging technique known as fingerprinting, which we will not cover in this course.

We saw how easy it can be to apply Theorem 1. All we need to do is select the right packing of databases, and the results follow easily. We will now see the same approach applied to proving lower bounds for histograms, though we also obtain a dependence on $\alpha$ as well. The same changes can be made to obtain a similar dependence for one-way marginals.

**Theorem 3.** *Any $\varepsilon$-DP algorithm $M : [k]^n \to [0,1]^k$ which estimates all histogram counts to error $\leq \alpha$ with probability $\geq 1/2$ requires $n = \Omega\left(\frac{\log k}{\alpha\varepsilon}\right)$.*

*Proof.* For a dataset $X \in [k]^n$, let $h(X) \in [0,1]^k$ be its (normalized) histogram representation. Let $y(X) \subset [0,1]^k$ be the $\ell_\infty$-ball of radius $\alpha$ around $h(X)$ – the accuracy guarantee says that $\Pr[M(X) \in y(X)] \geq 1/2$ for all databases $X$. For Theorem 1 to apply, we require that these sets are disjoint for a constructed set of databases, we specify this now. We construct $k$ databases $D_1$ through $D_k$: for each $\ell \in [k]$, database $D_\ell$ will have $t = 2\alpha n$ copies of $\ell$, and $(1 - 2\alpha)n$ copies of 1. Letting $Y_\ell = y(D_\ell)$, it is not hard to see that these sets are disjoint – consider $Y_\ell$'s inflated values on coordinate $\ell$ in comparison to the other sets. Observe that all databases can be converted to $D_1$ by changing at most $t$ points, since there are at most $t$ points which are not 1. Applying Theorem 1 gives the following bound:

$$\frac{1}{k} \geq e^{-\varepsilon t}/2.$$

Taking the logarithm of both sides and rearranging, we have $t \geq \frac{\log(k/2)}{\varepsilon}$, or $n = \Omega\left(\frac{\log k}{\alpha\varepsilon}\right)$. □