My notes for Grahams writeup

Ari

July 23, 2021

1 Lemma 1

The primary idea is to setup the idea of a bad event and show that its rare. I have come accross this style of proof technique a lot in Bandits, (example page 7 uniform exploration of `https://arxiv.org/pdf/1904.07272.pdf` proof). Once you can set up a bad event, you can put all the bad events in the δ bit of the Differential privacy definition – and deal with the good event. The good events are when approximate means are well behaved and are within the concentration inequalities of the expected count.

By assuming sampling with replacement, the odds of drawing a prefix remain the same in every single one of the m samples. the odds of a particular prefix being selected becomes Bernoulli (p) where $p = \frac{k}{n}$; k being the total counts of that prefix in D and n is cardinality of mulit set D. At this point all that remains is to bound the approximate sum behaviour with Chernoff Bound.

Based on my understanding the reason for the case by case is clear – we can't just kill of the $E[X]$ in the multiplicative bound, unless we are guaranteed to have it greater than 1. There might a slight issue with Case II however. There are few work arounds, change from $\theta E[X]$ to $(\theta + 1)E[X]$. But this will affect the additive bounds derivation.

Case I: $p \leq \frac{1}{m}$ or $mp \leq 1$ i.e. $E[X] \leq 1$. the expected number of times we pick this prefix is less than equal to 1. With $\theta > 1$, we have $E[X]\theta \leq \theta$, thus if $P\Big(X \geq \theta\Big)$, it implies $P\Big(X \geq E[X]\theta\Big)$. So the additive bound gives us the multiplicative bound.

$$P\Big(X \geq \theta\Big) = P\Big(\frac{X}{m} - \frac{E[X]}{m} \geq \frac{\theta}{m} - \frac{E[X]}{m}\Big)$$
$$= P\Big(\frac{X}{m} - \frac{mp}{m} \geq \frac{\theta}{m} - \frac{mp}{m}\Big)$$
$$= P\Big(\frac{X}{m} - p \geq \frac{\theta}{m} - p\Big)$$

By the additive Chernoff Bound

$$\leq exp\Big(- D(\frac{\theta}{m}||p)m\Big)$$

$$-D(\frac{\theta}{m}||p)m = -\theta ln\frac{\theta}{mp} - (m-\theta)ln\big(\frac{(m-\theta)/m}{1-p}\big)$$

Since $p \leq \dfrac{1}{m}$

$$-D(\frac{\theta}{m}||p)m- \leq \theta ln\frac{\theta m}{m} - (m-\theta)ln\big(\frac{(m-\theta)/m}{(m-1)/m}\big)$$
$$= -\theta ln\theta - (m-\theta)ln\big(\frac{m-\theta}{m-1}\big)$$
$$= -\theta ln\theta + (m-\theta)ln\big(\frac{m-1}{m-\theta}\big)$$
$$= -\theta ln\theta + (m-\theta)ln\big(1 + \frac{\theta-1}{m-\theta}\big)$$

This bit is a little magic to me

I do not have intution for this step

$$\leq -\theta ln\theta + (\theta - 1)$$

Case II: $p > \frac{1}{m}$, I do not think we can use this version of the Chernoff Bound. From the wikipedia page, where X is as defined by us, sum of Bernoulli random variables,

$$P\Big(X \geq (1+\delta)E[X]\Big) \leq e^{-\frac{\delta^2 E[X]}{3}}$$

but this requires $0 \leq \delta \leq 1$, therefore $1 \leq 1+\delta \leq 2$. In our case $\theta = (1+\delta) > 1$ is the only restriction. I think we need to use

$$P\Big(X \geq (1+\delta)E[X]\Big) \leq e^{-\frac{\delta^2 E[X]}{2+\delta}}$$

which requires $\delta \geq 0$; if we are willing change to bound to $(\theta+1)E[X]$, then $\delta = \theta$, now the bound still comes to the same.

$$\frac{\theta^2 E[X]}{2+\theta} > \frac{\theta^2}{2+\theta}$$
$$\text{Since} E[X] > 1$$

Assume: Set $\theta' = \theta + 1$; We get

$$P\Big(X \geq E[X](\theta+1)\Big) \leq e^{-\frac{(\theta'-1)^2}{1+\theta'}}$$

The new bound does not effect the final proof where the additive bound is still much more loose than the new bound. So the rest is fine.

$$\begin{aligned} P\Big(X \geq \theta'\Big) &\leq e^{-\theta' ln \theta' + (\theta'-1)} \\ &= e^{-(\theta+1) ln(\theta+1)+\theta} \end{aligned}$$

2 Lemma 2

Small typo I pointed out in the email. The rest is easy to understand – just the combinatorial expansion, identical to the original paper.

3 Lemma 3

Suppose at round i, the count for t_i, denoted as k, is less than $\frac{n}{m}$. Note we are only considering "Good events" E, where the counts are well behaved by Lemma 1. If we assume that the total number of prefixes for t_i in the data set D is k and $k < \frac{n}{m}$. This implies that $(\theta+1)E[\tau] = (\theta+1)\frac{k}{n/m} < (\theta+1)$, thereby the prefix will never be selected under a good event, where τ is the number of samples picked.

The other case is where $k \geq \frac{n}{m}$; where under a good event the number of t_i samples τ is at most $\frac{(\theta+1)m(k+1)}{n}$

$$\frac{P[M(D) \in R|E]}{P[M(D') \in R|E]} = \frac{k+1}{k+1-\tau}$$

Comes from Lemma 2 directly

$$\leq \frac{k+1}{k+1-\frac{(\theta+1)m(k+1)}{n}}$$

Substituting upper bound for τ

$$= \frac{n(k+1)}{n(k+1)-(\theta+1)m(k+1)}$$
$$= \frac{n}{n-(\theta+1)m}$$
$$= \frac{n}{n-\theta' m}$$
$$= \frac{n-\theta' m+\theta' m}{n-\theta' m}$$
$$= 1+\frac{\theta' m}{n-\theta' m}$$

where $\theta' = \theta + 1$ and assume $m \leq \frac{n}{10\theta'}$, then $\theta' m \leq \frac{n}{10}$ so

$$\frac{\theta' m}{n-\theta' m} \leq \frac{\theta' m}{n-n/10}$$
$$= \frac{10\theta' m}{9n}$$

The next step is clear. As each interactive round is independent and goes on for at the most L rounds, the final probability over all rounds is just the products. And the same bound is just exponentiated to L.

Side note: I did not really understand the following sentence, but I do not think it affects the proofs in large. Any string $t' \neq t$, if $t'_i \neq t_i$, then at level i, t'_i has the same number of copies in D' and D.

The last bound comes from the law of total probability

$$\begin{aligned}
P[M(D) \in R] &= P[M(D) \in R, E] + P[M(D) \in R, E^c] \\
&= P[E]P[M(D) \in R|E] + P[E^c]P[M(D) \in R|E^c] \\
&\leq P[E]P[M(D) \in R|E] + P[E^c] \\
\text{Since } P[M(D) \in R|E^c] \leq 1 & \\
&= P[E] \sum_{T \in R} P[M(D) = T \in R|E] + P[E^c] \\
&= P[E]\exp(\epsilon) \sum_{T \in R} P[M(D') = T \in R|E] + P[E^c] \\
&= P[E]\exp(\epsilon)P[M(D') \in R|E] + P[E^c] \\
\text{Since } P[E] \leq 1 & \\
&\leq \exp(\epsilon)P[M(D') \in R|E] + P[E^c]
\end{aligned}$$

4 Accuracy Bounds

Directly comes from the multiplicative bound Chernoff Bound.

$$\begin{aligned}
P\Big[X \leq \theta\Big] &= P\Big[X \leq \frac{w}{w}\theta\Big] \\
&= P\Big[X \leq (1 - \frac{w-\theta}{w})E[X]\Big] \\
\text{Set } \delta = \frac{w-\theta}{w} & \\
&= P\Big[X \leq (1-\delta)E[X]\Big] \\
&\leq e^{-\frac{\delta^2 E[X]}{2}} \\
&= e^{-\frac{(w-\theta)^2 w}{2w^2}} \\
&= e^{-\frac{(w-\theta)^2}{2w}}
\end{aligned}$$

5 An alternate view to TrieHH

Consider a specific round of the trie heavy hitter algorithm, say round l. In this round users vote on whether their prefix of size l can extend the trie. Let A represent the set of a size 1 prefixes. Then the set A^l represents all possible prefixes that can be added to the trie. For a trie of height l, each leaf node represents one element from A^l. Thus each row of the trie can be viewed as a histogram with $|A^l|$ bins/partitions. The figure below illustrates the idea. As the set of possible values is countable, we get write down a bijection from $f : A^l \to \mathbb{N}$.

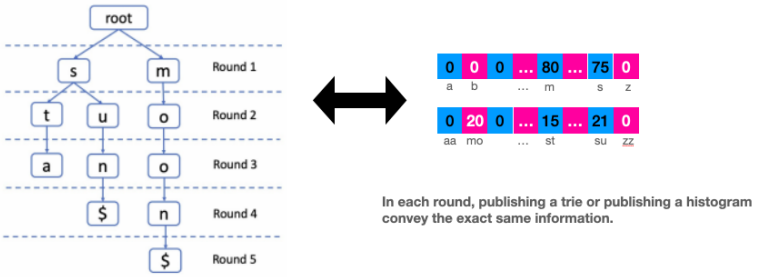

Figure 1: Example run of Algorithm 1.

Figure 1: At every round, the original protocol adds prefixes to its trie. This is equivalent to outputting a succinct histogram over the space of growing prefixes

If we publish the counts at each intermediate node, then we get a succinct histogram where only elements that occur more than θ times have a non zero probability of being added to the histogram. If we do not publish the counts, we see a 1 hot encoding of the same histogram with the same partitions. From here on we will describe the trie heavy hitter algorithm as publishing histograms at every round.

5.1 Sums and means are also differentially private

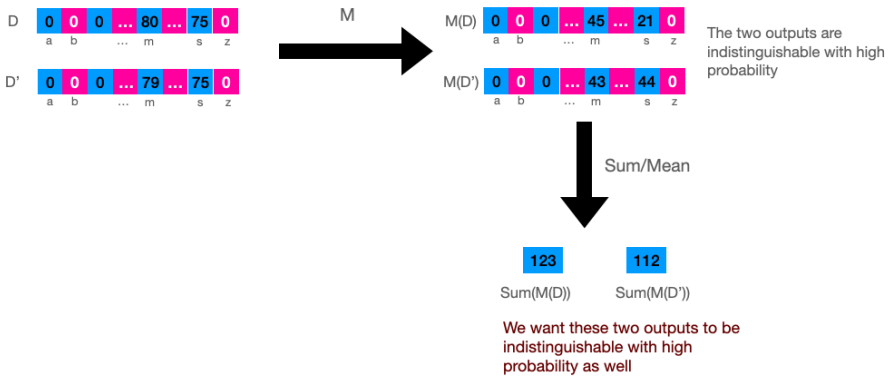

Figure 2: We would like to show that if we published the mean or sum of a subset of cells in the histogram, we would still preserve the privacy property.

So far we have shown that our algorithm M that publishes histograms $r \in R$, is differentially private i.e. $Pr\Big[M(D) \in r\Big] \leq exp(\epsilon)Pr\Big[M(D') \in r\Big] + \delta$. We

now show that, if instead of the histogram, we published sums or means of the histogram, the output is still differentially private. This is somewhat intuitive: If we can publish the full histogram without compromising differential privacy, outputting just the sum and mean which exposes less information, should not compromise differential privacy. See illustration in Figure 2

For our proof we consider histograms where the bins have an ordering order or cases like the alphabet where the bins have no strict ordering.

Consider two adjacet histograms D and D' shown above. Let them differ by one record at index k. If the input space consists of non integers like letters shown above, we can always learn a bijection of the input to the natural numbers. In the diagram above, cell m refers to positon k. Let d denote the number of bins in each histograms.

After application of M on the two histograms, under a good event E we get from Lemma 3:

$$\frac{Pr\Big[M_k(D) = \tau | E\Big]}{Pr\Big[M_k(D') = \tau | E\Big]} \leq 1 + \frac{10m\theta}{9n} = e^{\epsilon}$$

and

$$\frac{Pr\Big[M_j(D) = \tau\Big]}{Pr\Big[M_j(D') = \tau\Big]} = 1$$

for all $j \neq k$

We have

$$\begin{aligned} sum(M(D)) &= \sum_{i=1}^{d} M_i(D) * v_i \\ &= M_k(D) * v_k + \sum_{i \neq k} M_i(D) v_i \\ &= M_k(D) * v_k + X \end{aligned}$$

Similarly,

$$\begin{aligned} sum(M(D')) &= \sum_{i=1}^{d} M_i(D') * v_i \\ &= M_k(D') * v_k + \sum_{i \neq k} M_i(D') v_i \\ &= M_k(D') * v_k + X \end{aligned}$$

where v_i denote the weight of the i'th bin/index of the histogram. For our problem $v_i = 1 \ \forall i = \{1, .., d\}$. If users had integers instead of letters, then v_i would correspond the value of the integer. In the last step we introduce random

variable X to clean up our notation. It represents the sum of histogram bins that do not differ in D and D'.

$$\begin{aligned}
Pr\Big[sum(M(D)) = \tau | E\Big] &= \sum_{a,b|a+b=\tau} Pr\Big[X = b, M_k(D) * v_k = a | E\Big] \\
&= \sum_{a,b|a+b=\tau} Pr\Big[M_k(D) * v_k = a | E\Big] Pr\Big[X = b | E\Big] \\
&\leq \sum_{a,b|a+b=\tau} e^{\epsilon} Pr\Big[M_k(D') * v_k = a | E\Big] Pr\Big[X = b | E\Big] \\
&= e^{\epsilon} \sum_{a,b|a+b=\tau} Pr\Big[M_k(D') * v_k = a, X = b | E\Big] \\
&= e^{\epsilon} Pr\Big[sum(M(D')) = \tau | E\Big]
\end{aligned}$$

The second step comes from the cells of the histogram being independent. The third step comes from Lemma 3.

We can use the same logic as Grahams writeup to put all the non good events in the delta part of the definition.

$$\begin{aligned}
P[sum(M(D)) = \tau] &= P[sum(M(D)) = \tau, E] + P[sum(M(D)) = \tau, E^c] \\
&= P[E]P[sum(M(D)) = \tau | E] + P[E^c]P[sum(M(D)) = \tau | E^c] \\
&\leq P[E]P[sum(M(D)) = \tau | E] + P[E^c] \\
\text{Since } P[sum(M(D)) = \tau | E^c] \leq 1 & \\
&\leq P[E]\exp(\epsilon)P[sum(M(D')) = \tau | E] + P[E^c] \\
\text{Since } P[E] \leq 1 & \\
&\leq \exp(\epsilon)P[sum(M(D')) = \tau | E] + P[E^c]
\end{aligned}$$

The same argument would hold for means: it is just the sum scaled by the sampled population size m which is known before hand.

6 Relationship between additive noise and sub sampling

There is a lot of work about releasing histograms by adding Laplacian noise. The general idea is as follows: If we had two datasets D and D' that differed by 1 record only, we cannot just release the histograms in pure form. The difference of two histograms would identify the value of missing record which would break differential privacy. It can be shown that if one were to add Laplacian noise of scale $O(\frac{1}{\epsilon})$, then we get $(\epsilon, 0)$ differential privacy. However, in this scenario we add the same noise to each histogram cell regardless of the value in the cell. The noise only depends on the sensitivity of the query Δf, which for histogram

queries is 1. Thus the utility of each cell is affected the same way. In our world, we publish histograms and make them differentially private by sub sampling and killing off values below a certain threshold θ. For popular prefixes (i.e. high valued cells) our approximation is much closer to the true count than it would be for smaller valued cells. For cells that have values below θ we treat them all as 0.

Question: Is it worth analysing this further? There is so much work on the laplace mechanism but what are essentially saying is the same guarantees can be achieved by sub-sampling and thresholding.

References