A A TIGHTER BOUND FOR BINARY MECHANISM OF DP PREFIX-SUM

For a Laplace random variable, we have the following fact

Fact A.1 (Basic facts for Laplace random variable)**.** *Let x denote a random variable sampled from* $\mathsf{Lap}(b)$*, then*

$$\mathbb{E}[x^2] = 2b^2,$$

and with probability $1-\delta$, we have

$$|x| \le b\ln(1/\delta)$$

Proof. The CDF of $\mathsf{Lap}(b)$ is:

$$F(x) = \begin{cases} \frac{1}{2}\exp(x/b), & \text{if } x < 0; \\ 1 - \frac{1}{2}\exp(-x/b), & \text{if } x \ge 0. \end{cases}$$

we have:

$$F(b\ln(1/\delta)) = 1 - \delta/2.$$

This fact is proved by the symmetry of Laplace distribution. □

Lemma A.2. *Let $x_1, x_2, \ldots, x_n$ be n i.i.d. random variables sampled from Laplace distribution* $\mathsf{Lap}(b)$*. For $0 \le \delta \le 1$, with at most probability δ, we have for all $i \in [n]$:*

$$|x_i| > b\ln(n/\delta).$$

Proof. For Fact A.1, we know for each $i \in [n]$, with probability $p_i = \delta/n$:

$$|x_i| > b\ln(n/\delta)$$

Then, for all i, using Boole's inequality:

$$\Pr[|x_i| > b\ln(n/\delta)] \le \sum_{i=1}^{n} p_i = \delta.$$

□

Lemma A.3 (Bernstein inequality [17])**.** *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \le M$ almost surely (with probability $1-\delta$), for all i. Then, for all positive t,*

$$\Pr\left[\sum_{i=1}^{n} X_i > t\right] \le \exp\left(-\frac{t^2/2}{\sum_{j=1}^{n}\mathbb{E}[X_j^2] + Mt/3}\right) + \delta.$$

Lemma A.4 (similar to Lemma 2.8 in [24], sum of independent Laplace distributions)**.** *Let $x_1, x_2, \cdots, x_n$ denote n i.i.d. random variables sampled from Laplace distribution* $\mathsf{Lap}(b)$*. For all $t > 0$, we have*

$$\Pr\left[\sum_{i=1}^{n} x_i > t\right] \le \exp\Big(-\frac{t^2/2}{2nb^2 + Mt/3}\Big) + \delta$$

where $M = b\ln(n/\delta)$.

Proof. The proof is directly from Bernstein inequality and definition of Laplace distribution. From Lemma A.2, we have for all $i \in [n]$

$$|x_i| \le b\ln(n/\delta)$$

holds with at least probability $1-\delta$. □

Lemma A.5 (a slightly tighter version of Collary 2.9 in [24], measure concentration)**.** *Let $x_1, x_2, \ldots, x_n$ be n i.i.d. random variables sampled from Laplace distribution* $\mathsf{Lap}(b)$*. For any $0 \le \delta \le 1$, we have:*

$$\Pr[\sum_{i=1}^{n} x_i > \max\Big\{\sqrt{4nb^2\ln(3/\delta)}, (2/3)b\ln(3n/\delta)\cdot\ln(3/\delta)\Big\}] \le \delta$$

Proof. From Lemma A.4, we have for $0 \le \delta/3 \le 1$ ($M = b\ln(3n/\delta)$)

$$\Pr\left[\sum_{i=1}^{n} x_i > t\right] \le \exp\left(-\frac{t^2/2}{2nb^2 + b\ln(3n/\delta)\cdot t/3}\right) + \delta/3 = \exp\left(-\frac{t^2}{4nb^2 + (2/3)b\ln(3n/\delta)\cdot t}\right) + \delta/3$$

Now we prove an inequality for the general form of the exponent, for $a > 0, b > 0, t > 0, k > 0, t = \max\{\sqrt{ak}, bk\}$:

$$\frac{t^2}{c + bt} \ge \frac{t^2}{2\max\{a, bt\}} = \frac{\max\{ak, bkt\}}{2\max\{a, bt\}} = \frac{k}{2}.$$

We choose t as follows,

$$t = \max\{\sqrt{4nb^2\ln(3/\delta)}, (2/3)b\ln(3n/\delta)\ln(3/\delta)\},$$

then we have:

$$\Pr\left[\sum_{i=1}^{n} x_i > t\right] \le \exp(-0.5\ln(3/\delta)) + \delta/3 = \delta.$$

□

B DIFFERENTIALLY PRIVATE DISTINCT COUNT

In this section, we describe a differentially private distinct count algorithm based on [12]. We first prove the main technical lemmas about order statistics properties of random sampling in §4.1. Next, we define (ϵ, δ)-sensitivity and introduce the Laplacian mechanism for (ϵ, δ)-sensitivity in §4.2. This follows by our analysis of [12]: its (ϵ, δ)-sensitivity and its approximation ratio concentration. Last, we develop a differentially private distinct count algorithm based on [12] (Algorithm 8) and also demonstrate a simpler 1.1 approximation version (Algorithm 9).

B.1 Order Statistics Properties of Random Sampling

Claim B.1 (Restatement of Claim 4.1)**.** *Let $t \in [n]$, and fix any $0 < \delta < 1/2$. Then we have the following two bounds:*

(1) $\Pr[y_t > \delta\frac{t}{n}] \ge 1 - \delta$.
(2) $\Pr[y_t > \frac{t}{2n}] \ge 1 - \exp(-t/6)$.

Proof. **Part 1.** Consider the interval $[0, \delta\frac{t}{n}]$. We have

$$\mathbb{E}[|\{i \in [n] : x_i < \delta t/n\}|] = \delta t$$

namely, the expected number of points x_i will fall in this interval is exactly δt. Then by Markov's inequality, we have

$$\Pr[|\{i \in [n] : x_i < \delta t/n\}| \ge t] \le \delta.$$

So with probability $1-\delta$, we have $|\{i : x_i < \delta t/n\}| < t$, and conditioned on this we must have $y_t > \delta t/n$ by definition, which yields the first inequality.

Part 2. For the second inequality, note that we can write

$$|\{i \in [n] : x_i < t/(2n)\}| = \sum_{i=1}^{n} z_i$$

where $z_i \in \{0,1\}$ is an random variable that indicates the event that $x_i < t/2n$. Moreover, $\mathbb{E}[\sum_{i=1}^{n} z_i] = t/2$. Applying Chernoff bounds, we have

$$\Pr[|\{i \in [n] : x_i < t/(2n)\}| \geq t] \leq \exp(-t/6)$$

Which proves the second inequality. □

Claim B.2 (Restatement of Claim 4.2)**.** *Fix any $4 < \alpha < n/2$, and $1 \leq t \leq n/2$. Then we have*

$$\Pr[y_{t+1} < y_t + \alpha/n] \geq 1 - \exp(-\alpha/4).$$

Proof. Note that we can first condition on any realization of the values $y_1, y_2, \ldots, y_t$ one by one. Now that these values are fixed, the remaining distribution of the $(n-t)$ uniform variables is the same as drawing $(n-t)$ uniform random variables independently from the interval $[y_t, 1]$. Now observed that for any of the remaining $n-t$ uniform variables x_i, the probability that $x_i \in [y_t, y_t + \alpha/n]$ is at least $\frac{\alpha}{n}$, which follows from the fact that x_i is drawn uniformly from $[y_t, 1]$. Thus,

$$\begin{aligned}
&\Pr[|\{i \in S : x_i \in [y_t, y_t + \alpha/n]\}| = 0] \\
&\leq (1-\alpha/n)^{n-t} \\
&\leq (1-\alpha/n)^{n/2} \\
&= \exp(\frac{n}{2}\log(1-\alpha/n)) \\
&< \exp(\frac{n}{2}(-\alpha/n + 2(\alpha/n)^2)) \\
&< \exp(-\frac{n}{2}\frac{\alpha}{2n}) \\
&= \exp(-\alpha/4).
\end{aligned}$$

Thus $|\{i \in S : x_i \in [y_t, y_t + \alpha/n]\}| \geq 1$ with probability at least $1 - e^{-\alpha/4}$. Conditioned on this, we must have $y_{t+1} < y_t + \alpha/n$, as desired.

□

Lemma B.3 (Restatement of Lemma 4.3)**.** *Fix any $0 < \beta \leq 1/2$, $1 \leq t \leq n/2$, and α such that $4 < \alpha < \beta t/2$. Then we have the following two bounds:*

(1) $\Pr[|\frac{1}{y_t} - \frac{1}{y_{t+1}}| < \frac{\alpha}{\beta^2}\frac{n}{t^2}] \geq 1 - \beta - \exp(-\alpha/4)$.
(2) $\Pr[|\frac{1}{y_t} - \frac{1}{y_{t+1}}| < 4\alpha\frac{n}{t^2}] \geq 1 - \exp(-t/6) - \exp(-\alpha/4)$.

Proof. **Part 1.** For the first statement, we condition on $y_t > \beta\frac{t}{n}$ and, $y_{t+1} < y_t + \frac{\alpha}{n}$, which by a union bound hold together with probability $1 - \beta - e^{-\frac{\alpha}{4}}$ by Claims B.1 and B.2. Define the value t' such that $y_t = \frac{t'}{n}$. By the above conditioning, we know that $t' > \beta t$. Conditioned on this, we have

$$\begin{aligned}
\left|\frac{1}{y_t} - \frac{1}{y_{t+1}}\right| &< \frac{n}{t'} - \frac{1}{t'/n + \alpha/n} \\
&= \frac{n}{t'} - \frac{n}{t'+\alpha} \\
&= \frac{n}{t'}\left(1 - \frac{1}{1+\alpha/t'}\right) \\
&< \frac{n}{t'}\left(1 - (1-\alpha/t')\right) \\
&\leq \frac{\alpha n}{(t')^2} \\
&\leq \frac{\alpha n}{\beta^2 t^2} \qquad (1)
\end{aligned}$$

Where we used that $\alpha/t' < \alpha/(\beta t) < 1/2$, and the fact that $1/(1+x) > 1-x$ for any $x \in (0,1)$.

Part 2. For the second part, we condition on $y_t > \frac{t}{2n}$ and, $y_{t+1} < y_t + \frac{\alpha}{n}$, which by a union bound hold together with probability $1 - e^{-\frac{t}{6}} - e^{-\frac{\alpha}{4}}$ by Claims B.1 and B.2. Then from Lemma 1:

$$\begin{aligned}
\left|\frac{1}{y_t} - \frac{1}{y_{t+1}}\right| &\leq \frac{\alpha n}{\beta^2 t^2} \\
&= 4\alpha\frac{n}{t^2}
\end{aligned}$$

In this case, the same inequality goes throguh above with the setting $\beta = 1/2$, which finishes the proof. □

B.2 (ϵ, δ)-Sensitivity

In what follows, let $\mathcal{X}$ be the set of databases, and say that two databases $X, X' \in \mathcal{X}$ are neighbors if $\|X - X'\|_1 \leq 1$.

Definition B.4. Let $f : \mathcal{X} \to \mathbb{R}$ be a function. We say that f has sensitivity ℓ if for every two neighboring databases $X, X' \in \mathcal{X}$, we have $|f(X) - f(X')| \leq \ell$.

Theorem B.5 (The Laplace Mechanism [35])**.** *Let $f : \mathcal{X} \to \mathbb{R}$ be a function that is ℓ-sensitive. Then the algorithm A that on input X outputs $A(X) = f(X) + Lap(0, \ell/\epsilon)$ preserves $(\epsilon, 0)$-differential privacy.*

In other words, we have $\Pr[A(X) \in S] = (1 \pm \epsilon)\Pr[A(X') \in S]$ for any subset S of outputs and neighboring data-sets $X, X' \in \mathcal{X}$. Now consider the following definition.

Definition B.6 ((ℓ, δ)-sensitive)**.** Fix a randomized algorithm $\mathcal{A} : \mathcal{X} \times R \to \mathbb{R}$ which takes a database $X \in \mathcal{X}$ and a random string $r \in R$, where $R = \{0,1\}^m$ and m is the number of random bits used. We say that $\mathcal{A}$ is (ℓ, δ)-sensitive if for every $X \in \mathcal{X}$ there is a subset $R_X \subset R$ with $|R_X| > (1-\delta)|R|$ such that for any neighboring datasets $X, X' \in \mathcal{X}$ and any $r \in R_X$ we have $|\mathcal{A}(X, r) - \mathcal{A}(X', r)| \leq \ell$

Notice that our algorithm for count-distinct is $(O(\alpha\frac{n}{t}), O(e^{-t} + e^{-\alpha}))$-sensitive, following from the technical lemmas proved above. We now show that this property is enough to satisfy (ϵ, δ)-differential privacy after using the Laplacian mechanism.

Lemma B.7. *Fix a randomized algorithm $\mathcal{A} : \mathcal{X} \times R \to \mathbb{R}$ that is (ℓ, δ)-sensitive. Then consider the randomized laplace mechanism $\overline{\mathcal{A}}$ which on input X outputs $\mathcal{A}(X, r) + Lap(0, \ell/\epsilon)$ where $r \sim R$ is uniformly random string. Then the algorithm $\overline{\mathcal{A}}$ is $(\epsilon, 2(1+e^{\epsilon})\delta)$-differentially private.*

Proof. Fix any neighboring datasets $X, X' \in \mathcal{X}$. Let $R^* = R_X \cap R_{X'}$ where $R_X, R_{X'}$ are in Definition B.6. Since $|R_X| > (1-\delta)|R|$ and $|R_{X'}| > (1-\delta)|R|$, we have $|R_X \cap R_{X'}| > (1-2\delta)|R|$. Now fix any $r \in R^*$. By Definition B.6, we know that $|\mathcal{A}(X, r) - \mathcal{A}(X', r)| < \ell$.

From here, we follow the standard proof of correctness of the Laplacian mechanism by bounding the ratio

$$\frac{\Pr[\mathcal{A}(X, r) + \mathrm{Lap}(0, \frac{\ell}{\epsilon}) = z]}{\Pr[\mathcal{A}(X', r) + \mathrm{Lap}(0, \frac{\ell}{\epsilon}) = z]}$$

for any $z \in \mathbb{R}$.

In what follows, set $b = \frac{\ell}{\epsilon}$

$$\begin{aligned}
&\frac{\Pr[\mathcal{A}(X, r) + \mathrm{Lap}(0, b) = z]}{\Pr[\mathcal{A}(X', r) + \mathrm{Lap}(0, b) = z]} \\
&= \frac{\Pr[\mathrm{Lap}(0, b) = z - \mathcal{A}(X, r)]}{\Pr[\mathrm{Lap}(0, b) = z - \mathcal{A}(X', r)]} \\
&= \frac{\frac{1}{2b}\exp(-|z - \mathcal{A}(X, r)|/b)}{\frac{1}{2b}\exp(-|z - \mathcal{A}(X', r)|/b)} \\
&= \exp\left((|z - \mathcal{A}(X', r)| - |z - \mathcal{A}(X, r)|)/b\right) \\
&\le \exp\left(|\mathcal{A}(X, r) - \mathcal{A}(X', r)|/b\right) \\
&\le \exp(\ell/b) \\
&\le e^{\epsilon},
\end{aligned}$$

where the forth step follows from triangle inequality $|x| - |y| \le |x - y|$, the last step follows from $\ell/b = \epsilon$.

It follows that for any set $S \subset \mathbb{R}$ and any $r \in R^*$, we have

$$\begin{aligned}
&\Pr\left[\mathcal{A}(X, r) + \mathrm{Lap}(0, b) \in S\right] \\
&\le e^{\epsilon} \cdot \Pr\left[\mathcal{A}(X', r) + \mathrm{Lap}(0, b) \in S\right],
\end{aligned}$$

where the randomness is taken over the generation of the Laplacian random variable $\mathrm{Lap}(0, b)$. Since this holds for all $r \in R^*$, in particular it holds for a random choice of $r \in R^*$, thus we have

$$\begin{aligned}
&\Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X, r) + Z \in S\right] \\
&\le e^{\epsilon} \cdot \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X', r) + Z \in S\right] \qquad (2)
\end{aligned}$$

Now since $|R^*| \ge (1-2\delta)|R|$, by the law of total probability we have

$$\begin{aligned}
&\Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R}\left[\mathcal{A}(X, r) + Z \in S\right] \\
&= \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X, r) + Z \in S\right] \cdot \Pr[r \in R^*] \\
&\quad + \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R\setminus R^*}\left[\mathcal{A}(X, r) + Z \in S\right] \cdot \Pr[r \notin R^*] \\
&< \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X, r) + Z \in S\right] \\
&\quad + \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R\setminus R^*}\left[\mathcal{A}(X, r) + Z \in S\right] \cdot 2\delta \\
&\le \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X, r) + Z \in S\right] + 2\delta \qquad (3)
\end{aligned}$$

Similarly, it follows that

$$\begin{aligned}
&\Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R}\left[\mathcal{A}(X', r) + Z \in S\right] \\
&> \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X', r) + Z \in S\right](1 - 2\delta) \\
&\ge \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X', r) + Z \in S\right] - 2\delta \qquad (4)
\end{aligned}$$

where the last step follows from probability $\Pr[] \le 1$.

Combining Eq. (2), (3) and (4), we have

$$\begin{aligned}
&\Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R}[\mathcal{A}(X, r) + Z \in S] \\
&\le \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}[\mathcal{A}(X, r) + Z \in S] + 2\delta \\
&\le e^{\epsilon} \cdot \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R^*}\left[\mathcal{A}(X', r) + Z \in S\right] + 2\delta \\
&\le e^{\epsilon} \cdot \left(\Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R}\left[\mathcal{A}(X', r) + Z \in S\right] + 2\delta\right) + 2\delta
\end{aligned}$$

where the first step follows from Eq. (3), the second step follows Eq. (2), and the last step follows from Eq. (4).

Now recall that for the actual laplacian mechanism algorithm $\overline{\mathcal{A}}$, for any database X we have

$$\Pr[\overline{\mathcal{A}}(X) \in S] = \Pr_{Z\sim\mathrm{Lap}(0,b), r\sim R}[\mathcal{A}(X, r) + Z \in S],$$

which complets the proof that $\overline{\mathcal{A}}$ is $(\epsilon, 2(1 + e^{\epsilon})\delta)$-differentially private.

□

B.3 Analysis of Distinct Count

In this section, we thoroughly analyze the properties of Distinct Count [12]. We first describe the algorithm in Algorithm 7. Then we prove its (ℓ, δ)-sensitivity and a tighter (ϵ, δ)-approximation result (compared with the approximation result in [12]).

Algorithm 7 Distinct Count [12]

```
1:  procedure DistinctCount(I, t)                          ▷ Lemma B.8
2:      d ← ∅                                  ▷ d is a priority-queue of size t
3:      for x_i ∈ I do
4:          y ← h(x_i)                          ▷ h: [m] → [0, 1], is a PRF
5:          if |d| < t then
6:              d.push(y)
7:          else if y < d.top() ∧ y ∉ d then
8:              d.pop()
9:              d.push(y)
10:         end if
11:     end for
12:     v ← d.top()
13:     return t/v
14: end procedure
```

Sensitivity of distinct count

Lemma B.8 (Sensitivity of DistinctCount). *Assume $r \in R$ is the source of randomness of the PRF in DistinctCount (Algorithm 7), where $R \in \{0,1\}^m$, n is the number of distinct element of the input, for any $16 < t < n/2$, DistinctCount is $(20\log(4/\delta)\frac{n}{t}, \delta)$-sensitive.*

Proof. We denote DistinctCount (Algorithm 7) $F : \mathcal{X} \times R \to \mathbb{R}$, and define the same y_t as Section 4.1. Thus, for two neighboring database $X, X' \in \mathcal{X}$ ($\|X\|_0 = n$):

$$|F(X,r) - F(X',r)| \le \max\left\{\left|\frac{t}{y_t} - \frac{t}{y_{t-1}}\right|, \left|\frac{t}{y_t} - \frac{t}{y_{t+1}}\right|\right\}.$$

Part 1. From second inequality of Lemma 4.3 (the case $\beta = 1/2$), for any $5 < t \le n/2$ and $4 < \alpha < t/4$, we have:

$$\Pr\left[\left|\frac{1}{y_t} - \frac{1}{y_{t+1}}\right| \le 4\alpha\frac{n}{t^2}\right] \ge 1 - \exp(-t/6) - \exp(-\alpha/4)$$

It follows that

$$\begin{aligned}
&\Pr\left[\left|\frac{t}{y_t} - \frac{t}{y_{t+1}}\right| \le 5\alpha\frac{n}{t}\right] \\
>\ &\Pr\left[\left|\frac{t}{y_t} - \frac{t}{y_{t+1}}\right| \le 4\alpha\frac{n}{t}\right] \\
=\ &\Pr\left[\left|\frac{1}{y_t} - \frac{1}{y_{t+1}}\right| \le 4\alpha\frac{n}{t^2}\right] \\
&\ge 1 - \exp(-t/6) - \exp(-\alpha/4) \\
&\ge 1 - \exp(-(t/4)\cdot(2/3)) - \exp(-\alpha/4) \\
&\ge 1 - \exp(-2\alpha/3) - \exp(-\alpha/4) \\
&\ge 1 - 2\exp(-\alpha/4) \qquad (5)
\end{aligned}$$

Set $\alpha = 4\log(4/\delta)$ in Eq. (5):

$$\Pr\left[\left|\frac{t}{y_t} - \frac{t}{y_{t+1}}\right| \le 20\log(4/\delta)\frac{n}{t}\right] \ge 1 - \delta/2$$

Part 2. Similarly, from the the second inequality of Lemma 4.3 (the case $\beta = 1/2$), for any $10 < t \le n/2$ and $4 < \alpha < t/4$, we have:

$$\begin{aligned}
&\Pr\left[\left|\frac{1}{y_{t-1}} - \frac{1}{y_t}\right| \le 4\alpha\frac{n}{(t-1)^2}\right] \\
&\ge 1 - \exp(-(t-1)/6) - \exp(-\alpha/4)
\end{aligned}$$

From $t > 16$, we know $0.8t^2 < (t-1)^2$ and $t - 1 > 0.75t > 0$. Thus:

$$\begin{aligned}
&\Pr\left[\left|\frac{t}{y_t} - \frac{t}{y_{t-1}}\right| \le 5\alpha\frac{n}{t}\right] \\
=\ &\Pr\left[\left|\frac{t}{y_t} - \frac{t}{y_{t-1}}\right| \le 4\alpha\frac{nt}{0.8t^2}\right] \\
>\ &\Pr\left[\left|\frac{t}{y_t} - \frac{t}{y_{t-1}}\right| \le 4\alpha\frac{nt}{(t-1)^2}\right] \\
=\ &\Pr\left[\left|\frac{1}{y_{t-1}} - \frac{1}{y_t}\right| \le 4\alpha\frac{n}{(t-1)^2}\right] \\
&\ge 1 - \exp(-(t-1)/6) - \exp(-\alpha/4) \\
&\ge 1 - \exp(-0.75t/6) - \exp(-\alpha/4) \\
&\ge 1 - \exp(-\alpha/2) - \exp(-\alpha/4) \\
&\ge 1 - 2\exp(-\alpha/4) \qquad (6)
\end{aligned}$$

Set $\alpha = 4\log(4/\delta)$ in Eq. (6):

$$\Pr\left[\left|\frac{t}{y_t} - \frac{t}{y_{t+1}}\right| \le 20\log(4/\delta)\cdot\frac{n}{t}\right] \ge 1 - \delta/2$$

Part 3. Now apply union bound combining the results of **Part 1.** and **Part 2.**. Hence, for any X, $0 < \delta < 1$, $16 < t < n/2$:

$$\Pr\left[|F(X,r) - F(X',r)| \le 20\log(4/\delta)\cdot\frac{n}{t}\right] \le 1 - \delta.$$

Now, we proved the sensitivity of Algorithm 7. □

Lemma for approximation guarantees

Lemma B.9 (Restatement of Lemma 4.10)**.** *Let $x_1, x_2, \ldots, x_n \sim [0,1]$ be uniform random variables, and let $y_1, y_2, \ldots, y_n$ be their order statistics; namely, y_i is the i-th smallest value in $\{x_j\}_{j=1}^n$. Fix $\eta \in (0, 1/2), \delta \in (0, 1/2)$. Then if $t > 3(1+\eta)\eta^{-2}\log(2/\delta)$, with probability $1 - \delta$ we have*

$$(1-\eta)\cdot n \le \frac{t}{y_t} \le (1+\eta)\cdot n.$$

Proof. We define I_1 and I_2 as follows

$$I_1 = [0, \frac{t}{n(1+\eta)}],\ I_2 = [0, \frac{t}{n(1-\eta)}].$$

First note that if $x \sim [0,1]$, $\Pr[x \in I_1] = \frac{t}{n(1+\eta)}$. Since we have n independent trials, setting $Z = |\{x_i : i \in I_1\}|$ we have $\mathbb{E}[Z] = \frac{t}{(1+\eta)}$.

Then by the upper Chernoff bound, we have

$$\Pr[Z > t] \le \exp\left(-\frac{\eta^2 t}{3(1+\eta)}\right) \le 1 - \delta/2.$$

Similarly, setting $Z' = |\{x_i : i \in I_2\}|$, we have $\mathbb{E}[Z'] = \frac{t}{(1-\eta)}$, so by the lower Chernoff bound, we have

$$\Pr[Z' < t] \le \exp(-\eta^2 t/2) \le 1 - \delta/2.$$

Thus by a union bound, we have both that $Z < t$ and $Z' > t$ with probability $1 - \delta$. Conditioned on these two events, it follows that $y_t \notin I_1$ but $y_t \in I_2$, which implies that $\frac{t}{n(1+\eta)} < y_t < \frac{t}{n(1-\eta)}$, and so we have

$$(1-\eta)n < \frac{t}{y_t} < (1+\eta)n$$

as desired.

□

B.4 Differentially Private Distinct Count

Theorem B.10 (main result)**.** *For any $0 < \epsilon < 1$, $0 < \eta < 1/2$, $0 < \delta < 1/2$, there is an distinct count algorithm (Algorithm 8) such that:*

1. *The algorithm is (ϵ, δ)-differentially private.*
2. *With probability at least $1 - \delta$, the estimated distinct count $\widetilde{A}$ satisfies:*

$$n \le \widetilde{A} \le (1+\eta)\cdot n,$$

where n is the number of distinct elements in the data stream.

The space used by the distinct count algorithm is

$$O\Big((\eta^{-2} + \epsilon^{-1}\eta^{-1}\log(1/\delta))\cdot\log(1/\delta)\cdot\log n\Big)$$

bits.

Algorithm 8 DPDistinctCount: Differentially Private Distinct Count

procedure DPDistinctCount($I, \epsilon, \eta, \delta$) ▷ Theorem B.10
  pQueue $\leftarrow \emptyset$ ▷ pQueue is a priority-queue of size t ▷ $t \geq \max\left(3(1+\eta/4)(\eta/4)^{-2} \cdot \log(6/\delta),\ 20\epsilon^{-1}(\eta/4)^{-1} \cdot \log(24(1+e^{-\epsilon})/\delta) \cdot \log(3/\delta)\right)$
  for $x_i \in I$ **do**
    $y \leftarrow h(x_i)$ ▷ h: $[m] \to [0,1]$, is a PRF
    if $|\text{pQueue}| < t$ **then**
      pQueue.push(y)
    else if $y <$ pQueue.top() $\wedge\ y \notin$ pQueue **then**
      pQueue.pop()
      pQueue.push(y)
    end if
  end for
  $v \leftarrow$ pQueue.top()
  ct $\leftarrow (1+\frac{3}{4}\eta)\frac{t}{v} + \text{Lap}(20\epsilon^{-1}\frac{n}{t}\log(24(1+e^{-\epsilon})/\delta))$
  return ct
end procedure

Proof. Let $\widetilde{F}_0$ be the result output by the original algorithm (Algorithm 7) with the same t. Our differentially private distinct count algorithm (Algorithm 8) essentially output $\widetilde{A} = (1+\frac{3}{4}\eta)\widetilde{F}_0 + \text{Lap}(\ell/\epsilon)$, where:

$$\ell = 20\frac{n}{t}\log(24(1+e^{-\epsilon})/\delta)$$
$$t = \max\Big\{3(1+\eta/4)(\eta/4)^{-2}\log(6/\delta),\ 20\epsilon^{-1}(\eta/4)^{-1}\cdot\log(24(1+e^{-\epsilon})/\delta)\cdot\log(3/\delta)\Big\}$$

From Lemma B.8, we know that for any $16 < t < n/2$, the original distinct count algorithm (Algorithm 7)

$$\left(20\log(4/\delta)\cdot\frac{n}{t},\ \delta\right) - \text{sensitive.}$$

After rescaling δ by a constant factor, it can be rewritten as

$$\left(20\log(24(1+e^{-\epsilon})/\delta)\cdot\frac{n}{t},\ \frac{\delta}{6(1+e^{-\epsilon})}\right) - \text{sensitive.}$$

Then, from Lemma B.7, we know that it becomes $(\epsilon, \delta/3)$-DP by adding $\text{Lap}(\ell/\epsilon)$ noise when outputting the estimated count, where ℓ is as defined above, which completes the proof of the first part of the Theorem.

Next, from Lemma B.9 and the fact that $t \geq 3(1+\eta/4)\cdot(\eta/4)^{-2}\cdot\log(6/\delta)$, we know that $(1-\frac{\eta}{4})n < \widetilde{F}_0 < (1+\frac{\eta}{4})n$ with probability at least $1-\delta/3$.

Since $(1-\frac{1}{4}\eta)(1+\frac{3}{4}\eta) \leq 1+\frac{1}{4}\eta$ for any $0 < \eta \leq 1$, we get:

$$\Pr\left[(1+\frac{1}{4}\eta)\cdot n < (1+\frac{3}{4}\eta)\widetilde{F}_0 < (1+\frac{1}{2}\eta)\cdot n\right] \geq 1-\delta/3$$

Next, using the exponential $\Theta(e^{-x})$ tails of the Laplace distribution, and the fact that $\ell = \Omega(\log(1/\delta)\frac{n}{t})$ and $t = \Omega(\epsilon^{-1}\eta^{-1}\log^2(1/\delta))$, we have:

$$\Pr\left[|\text{Lap}(\ell/\epsilon)| > \frac{1}{4}\eta n\right] \leq \delta/3.$$

Conditioned on both event that

$$|\text{Lap}(\ell/\epsilon)| < \frac{1}{4}\eta n \text{ and}$$
$$(1+\frac{1}{4}\eta)\cdot n < (1+\frac{3}{4}\eta)\cdot\widetilde{F}_0 < (1+\frac{1}{2}\eta)\cdot n$$

which hold together with probability $1-\delta$ by a union bound, it follows that the estimate $\widetilde{A}$ of the algorithm indeed satisfies $n \leq \widetilde{A} \leq (1+\eta)n$, which completes the proof of the approximation guarantee in the second part of the Theorem. Finally, the space bound follows from the fact that the algorithm need only store the identities of the t smallest hashes in the data stream, which requires $O(t\log n)$ bits of space, yielding the bound as stated in the theorem after plugging in

$$t = \Theta\Big((\eta^{-2}+\epsilon^{-1}\eta^{-1}\log(1/\delta))\cdot\log(1/\delta)\Big).$$

Thus, we complete the proof. □

Algorithm 9 1.1-approx. DPDistinctCount (Restatement of Algorithm 2)

procedure 1.1-Approx. DPDC(I, ϵ, δ) ▷ Lemma B.12
  pQueue $\leftarrow \emptyset$ ▷ pQueue is a priority-queue of size t ▷ $t = 10^3\epsilon^{-1}\log(24(1+e^{-\epsilon})/\delta)\log(3/\delta))$
  for $x_i \in I$ **do**
    $y \leftarrow h(x_i)$ ▷ h: $[m] \to [0,1]$, is a PRF
    if $|\text{pQueue}| < t$ **then**
      pQueue.push(y)
    else if $y <$ pQueue.top() $\wedge\ y \notin$ pQueue **then**
      pQueue.pop()
      pQueue.push(y)
    end if
  end for
  $v \leftarrow$ pQueue.top()
  ct $\leftarrow 1.075\frac{t}{v} + \text{Lap}(0.02n/\log(3/\delta))$
  return ct
end procedure

Claim B.11. *For any $0 < \delta \leq 10^{-3}$, $0.1 \leq \eta < 1$ and $0 < \epsilon < 1$, then we have*

$$3(1+\eta/4)\cdot(\eta/4)^{-2}\cdot\log(6/\delta)$$
$$\leq 25\epsilon^{-1}(\eta/4)^{-1}\cdot\log(24(1+e^{-\epsilon})/\delta)\cdot\log(3/\delta).$$

Proof. From $0 < \delta < 1$, we know:

$$\log(6/\delta) \leq 2\log(3/\delta)$$

It follows:

$$\text{LHS} \leq 3\left[(1+\eta/4)\cdot(\eta/4)^{-1}\right]\cdot(\eta/4)^{-1}\cdot 2\log(3/\delta)$$
$$\leq 6(4/\eta+1)\cdot(\eta/4)^{-1}\cdot\log(3/\delta)$$

From $\eta \geq 0.1$, we know $4/\eta + 1 \leq 41$. From $\delta \leq 10^{-3}$, we also know $\log(24/\delta) \geq 10$. Thus:

$$\begin{aligned} LHS &\leq 246(\eta/4)^{-1} \cdot \log(3/\delta) \\ &\leq 25 \cdot 10 \cdot (\eta/4)^{-1} \cdot \log(3/\delta) \\ &\leq 25 \log(24/\delta) \cdot (\eta/4)^{-1} \cdot \log(3/\delta) \\ &\leq 25 \log(24(1+e^{-\epsilon})/\delta) \cdot (\eta/4)^{-1} \cdot \log(3/\delta) \\ &\leq 25\epsilon^{-1}(\eta/4)^{-1} \cdot \log(24(1+e^{-\epsilon})/\delta) \cdot \log(3/\delta) \end{aligned}$$

Now we completes the proof. □

Lemma B.12. *For any $0 < \epsilon < 1$, $0 < \delta \leq 10^{-3}$, there is an distinct count algorithm (Algorithm 9) such that:*

1. *The algorithm is (ϵ, δ)-differentially private.*
2. *With probability at least $1 - \delta$, the estimated distinct count $\widetilde{A}$ satisfies:*

$$n \leq \widetilde{A} \leq 1.1n,$$

where n is the number of distinct elements in the data stream.

The space used by the distinct count algorithm is

$$O\Big((100 + 10\epsilon^{-1}\log(1/\delta)) \cdot \log(1/\delta) \cdot \log n\Big)$$

bits.

Proof. This lemma directly follows Theorem B.10 by setting $\eta = 0.1$, $0 < \delta < 10^{-3}$ and:

$$\begin{aligned} t &= 25\epsilon^{-1}(\eta/4)^{-1} \cdot \log(24(1+e^{-\epsilon})/\delta) \cdot \log(3/\delta) \\ &= 10^3\epsilon^{-1}\log(24(1+e^{-\epsilon})/\delta) \cdot \log(3/\delta) \end{aligned}$$

Thus, the scale factor in the Lap distribution (line 14 in Algorithm 8) becomes:

$$\begin{aligned} & 20\epsilon^{-1}\frac{n}{t}\log(24(1+e^{-\epsilon})/\delta) \\ =\ & 20\epsilon^{-1}\frac{n\log(24(1+e^{-\epsilon})/\delta)}{1000\epsilon^{-1}\log(24(1+e^{-\epsilon})/\delta) \cdot \log(3/\delta)} \\ =\ & 0.02n/\log(3/\delta) \end{aligned}$$

□

C PROPERTIES OF BINOMIAL DISTRIBUTION

Fact C.1 (Tail bounds of binomial distribution)**.** *If $X \sim B(n, p)$, that is, X is a binomially distributed random variable, where n is the total number of experiment and p is the probability of each experiment getting a successful result, and $k \geq np$, then:*

$$\Pr[X \geq k] \leq \exp(-2n(1 - p - (n-k)/n)^2)$$

Proof. For $k \leq np$, from the lower tail of the CDF of binomial distribution $F(k; n, p) = \Pr[X \leq k]$, we use Hoeffding's inequality [47] to get a simple bound:

$$F(k; n, p) \leq \exp(-2n(p - k/n)^2)$$

For $k \geq np$, since $\Pr[X \geq k] = F(n-k; n, 1-p)$, we have:

$$\Pr[X \geq k] \leq \exp(-2n(1 - p - (n-k)/n)^2).$$

□

Lemma C.2. *If $X \sim B(n, p)$, that is, X is a binomially distributed random variable, where n is the total number of experiment and p is the probability of each experiment getting a successful result, then*

$$\Pr[X \geq np + \sqrt{0.5n\log(1/\delta)}] \leq \delta$$

Proof. From Fact C.1, let $k = np + \sqrt{0.5n\log(1/\delta)}$, we have:

$$\begin{aligned} & \Pr[X \geq np + \sqrt{0.5n\log(1/\delta)}] \\ \leq\ & \exp(-2n(1 - p - (n - (np + \sqrt{-0.5n\log(1/\delta)}))/n)^2) \\ =\ & \exp(-2n\frac{1}{2n}\log(1/\delta)) \\ =\ & \delta. \end{aligned}$$

□