CSIS0351/CSIS8601: Randomized Algorithms
Lecture 10: Private and continual releasing of statistics
Instructor: Hubert Chan
Date: 21 Nov 2011

These lecture notes are supplementary materials for the lectures. They are by no means substitutes for attending lectures or replacement for your own notes!

We shall consider an application of differential privacy to a streaming problem. The material of this note is based on the paper [CSS10].

1 Problem Definition

We consider the following problem. Assume that the input stream $\sigma \in \{0,1\}^{\mathbb{N}}$ is a sequence of bits. The bit $\sigma(t)$ at time $t \in \mathbb{N}$ denotes whether an event of interest occurred at time t. At every time step $t \in \mathbb{N}$, we want to know how many 1's have appeared in the stream by time t.

Definition 1.1 (Continual Counting Query) *Given a stream $\sigma \in \{0,1\}^{\mathbb{N}}$, the count for the stream is a mapping $c_\sigma : \mathbb{N} \to \mathbb{Z}$ such that for each $t \in \mathbb{N}$, $c_\sigma(t) := \sum_{i=1}^{t} \sigma(i)$.*

Definition 1.2 (Counting Mechanism) *A counting mechanism $\mathcal{M}$ takes a stream $\sigma \in \{0,1\}^{\mathbb{N}}$ and produces a (possibly randomized) mapping $\mathcal{M}(\sigma) : \mathbb{N} \to \mathbb{R}$. Moreover, for all $t \in \mathbb{N}$, $\mathcal{M}(\sigma)(t)$ is independent of all $\sigma(i)'s$ for $i > t$. We can also view $\mathcal{M}(\sigma)$ as a point in $\mathbb{R}^{\mathbb{N}}$.*

Remark 1.3 Observe that we can always make the counting mechanism return non-negative integers. If the mechanism returns a real number, we can round (deterministically) to the nearest non-negative integer.

Definition 1.4 (Time-bounded Mechanism) *A counting mechanism $\mathcal{M}$ is unbounded, if it accepts streams of indefinite lengths, i.e., given any stream σ, $\mathcal{M}(\sigma) \in \mathbb{R}^{\mathbb{N}}$. Given $T \in \mathbb{N}$, a mechanism $\mathcal{M}$ is T-bounded if it only accepts streams of lengths at most T and returns $\mathcal{M}(\sigma) \in \mathbb{R}^T$. In other words, the mechanism needs to know the value T in advance and only looks at the length T prefix of any given stream.*

Definition 1.5 (Utility) *A counting mechanism $\mathcal{M}$ is (λ, δ)-useful at time t, if for any stream σ, with probability at least $1 - \delta$, we have $|c_\sigma(t) - \mathcal{M}(\sigma)(t)| \leq \lambda$.*

Definition 1.6 (Differential Privacy) *Two streams σ and σ' are neighboring, if they differ at exactly one time t. Then, the definition of ϵ-differential privacy for counting mechanisms follows from the standard one, i.e., for any neighboring streams σ and σ', and any measurable subset $S \subseteq \mathbb{R}^{\mathbb{N}}$ (or $S \subseteq \mathbb{R}^T$ for T-bounded mechanisms), $\Pr[\mathcal{M}(\sigma) \in S] \leq \exp(\epsilon) \cdot \Pr[\mathcal{M}(\sigma') \in S]$.*

In this lecture, we give a general framework for constructing ϵ-differentially private counting mechanisms, and give several concrete constructions.

2 A Technical Lemma

In the our construction of differentially private counting mechanisms, the random noise might be a sum of more than one Laplace random variables. The following lemma states that sum of independent Laplace distributions is concentrated around 0. We leave its proof as an exercise.

Lemma 2.1 *Let $\gamma_1, \gamma_2, \ldots, \gamma_n$ be n independent random variables, where γ_i is sampled from $\mathsf{Lap}(b_i)$. Let $Y := \sum_{i=1}^{n} \gamma_i$. Suppose $0 < \delta < 1$. With probability at least $1 - \delta$, the random variable $|Y|$ is at most $O(\sqrt{\sum_{i=1}^{n} b_i^2} \cdot \log \frac{1}{\delta})$.*

Remark 2.2 In the counting problem, we could have used the geometric distribution for generating noise, since only integers are involved. There is a version of Lemma 2.1 for geometric distribution, but the proof is slightly more complicated.

3 Time Bounded Mechanisms

3.1 Simple Counting Mechanism I

In the time bounded case, a counting query can be viewed as a function $c : \{0,1\}^T \to \mathbb{R}^T$, where $c_i(\sigma) := c_\sigma(i)$. Suppose $\sigma \in \{0,1\}^T$ and $\sigma' \in \{0,1\}^T$ are two neighboring streams such that $\sigma(t) \neq \sigma'(t)$. Then we know that for $1 \leq i < t$, $c_\sigma(i) = c_{\sigma'}(i)$, and for $t \leq i \leq T$, $|c_\sigma(i) - c_{\sigma'}(i)| = 1$. This implies that the sensitivity of c is $\max_{\sigma \sim \sigma'} \sum_i |c_\sigma(i) - c_{\sigma'}(i)| \leq T$. Thus, we naturally get the following simple counting mechanism: at every time step t, we sample a fresh independent random variable $\gamma_t \sim \mathsf{Lap}(\frac{T}{\epsilon})$ and release $c(t) + \gamma_t$. By the properties of Laplace distribution, we know that this mechanism preservers ϵ-differential privacy; and by Lemma 2.1, the mechanism is $(O(\frac{T}{\epsilon} \log \frac{1}{\delta}), \delta)$-useful at each time $t \in [T]$.

3.2 Simple Counting Mechanism II

The Simple Counting Mechanism I generates the true output and add random noise to perturb it. Another natural way to achieve privacy is to first perturb the input, and then do calculations on top of it to get private and useful output.

We think the input as a function of itself and consider the function $I : \{0,1\}^T \to \mathbb{R}^T$ such that $I(\sigma) = \sigma$. For any two neighboring streams σ and σ' such that $\sigma(t) \neq \sigma'(t)$, we have that $I_i(\sigma) = \sigma(i) = \sigma'(i) = I_i(\sigma')$ for $i \neq t$, and that $I_t(\sigma) = \sigma(t) \neq \sigma'(t) = I_t(\sigma)$. Hence, the sensitivity of I is 1. Thus, the function $\widehat{I}$ such that $\widehat{I}_i(\sigma) := I_i(\sigma) + \gamma_i$, where $\gamma_1, \gamma_2, \ldots, \gamma_T$ are independent random variable sampled from $\mathsf{Lap}(\frac{1}{\epsilon})$, is ϵ-differentially private.

Note that for any stream $\sigma \in \{0,1\}^T$, we have $c_\sigma(t) = \sum_{i=1}^{t} I_i(\sigma)$. Hence, we can output $\widehat{c}_\sigma(t) := \sum_{i=1}^{t} \widehat{I}_i(\sigma)$ at time step t. Observe that $\widehat{c}$ is a deterministic function of $\widehat{I}$. Hence, $\widehat{c}$ also preserves ϵ-differential privacy.

Fix $t \in [T]$. Note that $\widehat{c}_\sigma(t) = \sum_{i=1}^{t} \widehat{I}_i(\sigma) = \sum_{i=1}^{t} (\widehat{I}(\sigma) + \gamma_i) = c_\sigma(t) + \sum_{i=1}^{t} \gamma_i$. Hence, by Lemma 2.1, we know that simple counting mechanism II is $(O(\frac{\sqrt{t}}{\epsilon} \log \frac{1}{\delta}), \delta)$-useful at time step t.

3.3 p-sum Framework

Simple Counting Mechanism II suggests a new idea of designing differentially private counting mechanisms with small error: although the counting query as a function has sensitivity as large as T, we can use functions with small sensitivity as ingredients to construct differentially private counting mechanisms.

In the following constructions, we consider partial sums as the ingredient.

Definition 3.1 (p-sum) *A p-sum is a partial sum of consecutive items. For $1 \leq i \leq j$, the p-sum $\Sigma[i,j] : \{0,1\}^T \to \mathbb{R}$ is a function such that $\Sigma[i,j](\sigma) = \sum_{k=i}^{j} \sigma(k)$.*

Let $\mathcal{C}$ be a collection of p-sums such that each count $c_\sigma(t)$ can be represented as a sum of p-sums involving only items arrived by time t, and that each item appears in at most α p-sums. Given an input stream $\sigma \in \{0,1\}^T$, the mechanism $\mathcal{M}(\mathcal{C}, \alpha)$ using p-sum framework works as follows.

- For each p-sum $\Sigma[i,j]$ in $\mathcal{C}$, add an independent Laplace random variable $\gamma_{ij} \sim \mathsf{Lap}(\frac{\alpha}{\epsilon})$ to it when it gets ready at time j, to get a noisy p-sum $\widehat{\Sigma}[i,j](\sigma) := \widehat{\Sigma}[i,j](\sigma) + \gamma_{ij}$.
- Let $t \in [T]$ be any time step. Suppose the interval $[1,t]$ can be covered by disjoint intervals $\mathcal{C}_t := \{[i_k, j_k]\}_k$. Then, $c_\sigma(t) = \sum_k \Sigma[i_k, j_k](\sigma)$. Then, at time t, every noisy p-sum $\widehat{\Sigma}[i_k, j_k]$ is ready, and we output $\widehat{c}_\sigma(t) := \sum_k \widehat{\Sigma}[i_k, j_k](\sigma)$ as the approximate count at time t.

For mechanisms using p-sum framework, we have the following theorem.

Theorem 3.2 *Let $\mathcal{C}$ be a collection of p-sums such that*

1. *Each item in σ appears in at most α p-sums in $\mathcal{C}$.*
2. *Each count $c_\sigma(t)$ is a sum of at most β p-sums in $\mathcal{C}$ which only involves items arriving by time t.*

Then, the mechanism $\mathcal{M}(\mathcal{C}, \alpha)$ using the above framework

- *preserves ϵ-differential privacy,*
- *is $(O(\frac{\alpha\sqrt{\beta}}{\epsilon} \cdot \log \frac{1}{\delta}), \delta)$-useful at each time t.*

Proof: Let σ and σ' be two neighboring streams such that $\sigma(t) \neq \sigma'(t)$. We view the p-sums in $\mathcal{C} = \{\Sigma[i_k, j_k]\}$ as coordinates of a vector function $f : \{0,1\}^T \to \mathbb{R}^{|\mathcal{C}|}$ such that $f_k(\sigma) = \Sigma[i_k, j_k](\sigma)$. Since each item appears in at most α p-sums, we know that there are at most α p-sums such that $|\Sigma[i,j](\sigma) - \Sigma[i,j](\sigma')| = 1$, and the remaining p-sums have the same value for σ and σ'. This implies that the function f has sensitivity at most α. Hence, we know that the noisy p-sums preserve ϵ-differential privacy. Also, the outputted approximate count $\widehat{c}_\sigma(t)$ is a deterministic function of the noisy p-sums. Hence, the mechanism preservers ϵ-differential privacy.

Let $t \in [T]$ be any time step, and suppose count $c_\sigma(t)$ can be represented as a sum of at most β p-sums, i.e., $c_\sigma(t) = \sum_k \Sigma[i_k, j_k](\sigma)$. We know that the additive error of the output $|\widehat{c}_\sigma(t) - c_\sigma(t)| =$

$|\sum_k \widehat{\Sigma}[i_k, j_k](\sigma) - \sum_k \Sigma[i_k, j_k](\sigma)| = |\sum_k \gamma_{i_k,j_k}|$. Note that each random noise γ_{i_k,j_k} is sampled from $\mathsf{Lap}(\frac{\alpha}{\epsilon})$. By Lemma 2.1, we know that with probability at least δ, the additive error is at most $O(\sqrt{\sum_k (\frac{\alpha}{\epsilon})^2} \log \frac{1}{\delta}) = O(\frac{\alpha\sqrt{\beta}}{\epsilon} \log \frac{1}{\delta})$. ■

Let us rethink Simple Counting Mechanism I and II using the p-sum framework. In Simple Counting Mechanism I, we use $\mathcal{C} := \{\Sigma[1,t] : \forall t \in [T]\}$ and decompose $c_\sigma(t)$ as only one p-sum $\Sigma[1,t](\sigma)$. Note that $\sigma(t)$ appears in at most T p-sums, hence we add random noise sampled from $\mathsf{Lap}(\frac{T}{\epsilon})$. Also, each count $c_\sigma(t)$ is represented by exactly one p-sum. Therefore, the mechanism is $(O(\frac{T}{\epsilon} \log \frac{1}{\delta}), \delta)$-useful at each time t.

In Simple Counting Mechanism II, we choose $\mathcal{C} := \{\Sigma[t,t] : \forall t \in [T]\}$, and decompose $c_\sigma(t)$ as $\sum_{i\in[t]} \Sigma[i,i](\sigma)$. Note that each item appears in exactly one p-sum. So we add random noise sampled from $\mathsf{Lap}(\frac{1}{\epsilon})$. Also, each count $c_\sigma(t)$ is represented by a sum of t p-sums. Hence, the mechanism is $(O(\frac{\sqrt{t}}{\epsilon} \cdot \log \frac{1}{\delta}), \delta)$-useful at time t.

3.4 Two-level Mechanism

Note that Theorem 3.2 implies that in order to improve the utility of private counting mechanisms using p-sum framework, we need to let α and β be as small as possible. Simple Counting Mechanism I has a small β with a sacrifice of large $\alpha = T$, while Simple Counting Mechanism II uses a β as large as T to trade in for a small $\alpha = 1$. The challenge is to strike a balance between the optimization for α and β.

The two-level mechanism uses the idea of grouping the items in the stream as blocks of size B. Note that the prefix $\sigma([t])$ with $t = qB + r$ is a union of q blocks together with r single items. Hence, we choose the following p-sums as $\mathcal{C}$: for each k such that $kB \le T$, we choose $\Sigma[(k-1) \cdot B + 1, kB]$; also, we include $\Sigma[i,i]$ for each $i \in [T]$. Then, each count $c_\sigma(t)$ can be represented as $\sum_{i=1}^{q} \Sigma[(i-1)B + 1, iB] + \sum_{j=1}^{r} \Sigma[qB + j, qB + j]$, where $t = qB + r$ and $0 \le r < B$. Note that $q \le \frac{t}{B}$ and $r \le B$. Hence, count $c_\sigma(t)$ is a sum of at most $\frac{t}{B} + B$ p-sums.

Let $t = qB + r$ be any time. Then, the only items that include $\sigma(t)$ is $\Sigma[qB + 1, (q+1)B]$ and $\Sigma[t,t]$. Hence, each item appears in at most two p-sums.

Using the p-sum framework, we get the Two-level Mechanism, $\mathcal{M}(\mathcal{C}, 2)$. By Theorem 3.2, it preserves ϵ-differential privacy and is $(O(\frac{2\sqrt{B + \frac{t}{B}}}{\epsilon} \cdot \log \frac{1}{\delta}), \delta)$-useful at time t. Note that if we choose $B := \left\lceil \sqrt{T} \right\rceil$, then we get a ϵ-differentially private counting mechanism that is $(O(\frac{T^{\frac{1}{4}}}{\epsilon} \cdot \log \frac{1}{\delta}), \delta)$-useful at every time t.

3.5 Binary Mechanism

The Binary Mechanism utilizes the fact that any integer t can be represented as a sum of at most $\lceil \log t \rceil$ powers of 2, and hence each prefix $\sigma(t)$ can be decomposed into at most $\lceil \log t \rceil$ blocks of sizes which are powers of 2. For each $0 \le \ell \le \lceil \log T \rceil$, we choose $\mathcal{C}_\ell := \{\Sigma[(k-1)2^i + 1, k2^i] : k2^i \le T\}$. Then, we choose $\mathcal{C} := \cup_{\ell=0}^{\lceil \log t \rceil} \mathcal{C}_\ell$.

Note that for each ℓ, the p-sums in $\mathcal{C}_\ell$ are disjoint. So each item $\sigma(t)$ appears in at most one p-sum in every $\mathcal{C}_\ell$, and hence appears in at most $\lceil \log T \rceil$ p-sums in $\mathcal{C}$.

For each t, let $\mathsf{Bin}_i(t) \in \{0,1\}$ be the i-th digit in the binary representation of t, where $\mathsf{Bin}_0(t)$ is the least significant digit. Hence, $t = \sum_i \mathsf{Bin}_i(t)2^i$. Note that for each i, $\sum_{j>i} \mathsf{Bin}_j(t)2^j = (\sum_{j>i} \mathsf{Bin}_j(t)2^{j-i}) \cdot 2^i$, which implies that the p-sum $[\sum_{j>i} \mathsf{Bin}_j(t)2^j + 1, \sum_{j>i} \mathsf{Bin}_j(t)2^j + 2^i] \in \mathcal{C}_i$. Also observe that $c_\sigma(t) = \sum_{i:\mathsf{Bin}_i(t)=1} \Sigma[\sum_{j>i} \mathsf{Bin}_j(t)2^j + 1, \sum_{j>i} \mathsf{Bin}_j(t)2^j + 2^i]$. Hence, each count $c_\sigma(t)$ can be represented as a sum of at most $\lceil \log T \rceil$ p-sums in $\mathcal{C}$.

By Theorem 3.2, we know that the mechanism $\mathcal{M}(\mathcal{C}, \lceil \log T \rceil)$ preserves ϵ-differential privacy, and is $(O(\frac{\log^{1.5} T}{\epsilon} \cdot \log \frac{1}{\delta}), \delta)$-useful at every time t.

4 Unbounded Counting Mechanisms

Note that Simple Counting Mechanism II does not require a prior knowledge about the time bound T. Hence, it is an unbounded counting mechanism. However, as t grows, the additive error at time t has a $\Theta(\sqrt{t})$ dependence on t.

Using binary mechanism as a building block, we can get an unbounded counting mechanism that is $(O(\frac{\log^{1.5} t}{\epsilon} \cdot \log \frac{1}{\delta}), \delta)$-useful at every time t. The details are included in the paper [CSS10].

5 Homework Preview

1. In this question, we derive a measure concentration result for independent random variables drawn from Laplace distribution. We show that with high probability, the sum of independent Laplace random variables are concentrated around its mean, 0.

   We use moment generating functions in a Chernoff-like argument. Let $\gamma_1, \gamma_2, \ldots, \gamma_n$ be n independent random variables, where γ_i is sampled from $\mathsf{Lap}(b_i)$.

   (a) Prove that for each γ_i, the moment generating function is $E[\exp(h\gamma_i)] = \frac{1}{1-h^2 b_i^2}$, where $|h| < \frac{1}{b_i}$.

   (b) Show that $E[\exp(h\gamma_i)] \leq \exp(2h^2 b_i^2)$, if $|h| < \frac{1}{\sqrt{2}b_i}$.
   (Hint: for $|x| < \frac{1}{2}$, we have $\frac{1}{1-x} \leq 1 + 2x \leq \exp(2x)$.)

   (c) Let $b_M := \max_{i\in[n]} b_i$. Also, let $\nu \geq \sqrt{\sum_{i=1}^n b_i^2}$ and $0 < \lambda < \frac{2\sqrt{2}\nu^2}{b_M}$. Prove that $\Pr[|Y| > \lambda] \leq 2\exp\left(-\frac{\lambda^2}{8\nu^2}\right)$

   (d) Suppose $0 < \delta < 1$ and $\nu > \max\left\{\sqrt{\sum_{i=1}^n b_i^2}, b_M\sqrt{\ln\frac{2}{\delta}}\right\}$. Prove that $\Pr[|Y| > \nu\sqrt{8\ln\frac{2}{\delta}}] \leq \delta$.

   (e) Prove that $\Pr[|Y| > \sqrt{8} \cdot \sqrt{\sum_{i=1}^n b_i^2} \cdot \ln\frac{2}{\delta}] \leq \delta$.

References

[CSS10] T.-H. Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. In *ICALP (2)*, pages 405–417, 2010.