ISYE/CS 8803: Foundations of Data Privacy **Fall 2017**

Lecture 5: Properties of Differential Privacy & Sparse Vector

Lecturer: Rachel Cummings *Scribe: Nayeon Kim, 9/6/2017*

1 Properties of Differential Privacy (DP)

We have seen three mechanisms for achieving differential privacy (DP): Laplace mechanism, Exponential mechanism, and Randomized response. Using those as building blocks, we can construct more complicated mechanisms and perform more advanced data analysis. Following properties of DP allow us to build a new mechanisms using the above mechanisms while maintaining their privacy guarantees.

Theorem 1. *(Post-processing) Let $M : \mathbb{N}^{|X|} \to R$ be (ϵ, δ)-DP and let $f : R \to R'$ be an arbitrary randomized function. Then, $f \circ M : \mathbb{N}^{|X|} \to R'$ is (ϵ, δ)-DP.*

Proof. Proposition 2.1 in text book [DR14] (p19) □

Theorem 1 guarantees that no additional information about the database will be learned by adversaries through further computations or process on a DP output. Notice that there is no assumption on the computational power of the adversary in Theorem 1.
DP is phrased in terms of a single individual, but in practice, we may want privacy for groups. For example, there can be families whose data is correlated or identical or we may consider multiple entries in the database of a same individual such as multiple hospital records of a same patient. Following theorem provides privacy of a group of data.

Theorem 2. *(Group Privacy) Let $M : \mathbb{N}^{|X|} \to R$ be (ϵ, δ)-DP. Then M is also $(k\epsilon, ke^{(k-1)\epsilon}\delta)$-DP for groups of size k. That is, for all x, y such that $||x - y||_1 \leq k$ and for all $S \subseteq R$,*

$$P[M(x) \in S] \leq e^{k\epsilon} P[M(y) \in S] + ke^{(k-1)\epsilon}\delta$$

Theorem 2 shows the level of privacy degrades linearly with a size of group.
Composition is another useful property of DC. The simple algorithms can be combined to build an advanced algorithm and their privacy guarantees degrade as multiple computations are performed on the same database.

Theorem 3. *(Basic Composition) Let $M_i : \mathbb{N}^{|X|} \to R_i$ be (ϵ_i, δ_i)-DP for $i = 1, \ldots, k$. Then the composition $M_{[k]} : \mathbb{N}^{|X|} \to R_1 \times \cdots \times R_k$ defined as:*

$$M_{[k]}(x) = (M_1(x), M_2(x), \ldots, M_k(x))$$

is $(\sum_{i=1}^{k} \epsilon_i, \sum_{i=1}^{k} \delta_i)$-DP.

Proof. Theorem 3.14 and Corollary 3.15 in text book [DR14] (p42) □

Theorem 3 indicates that, for example, when we run k mechanisms, each mechanism can be set to be $\frac{\epsilon}{k}$-DP to achieve overall privacy of ϵ-DP, where ϵ is a privacy budget.
There is also an advanced composition, which improves on basic composition in two ways. First, it allows ϵ to degrade in $\mathcal{O}(\sqrt{k}\epsilon)$ instead of $\mathcal{O}(k\epsilon)$ as in basic composition, at a small increase in δ. Second, it allows adaptive composition, where the choice of the i-th mechanism can depend on the previous i-1 mechanisms and their outputs. That is, the i-th mechanism can be written as:

$$M : \mathbb{N}^{|X|} \times (M_1 \times \cdots \times M_{i-1}) \times (R_1 \times \cdots \times R_{i-1})$$

Theorem 4. *(Advanced Composition) Let $M : \mathbb{N}^{|X|} \to R^k$ be a k-fold adaptive composition of (ϵ, δ)-DP mechanisms. Then M is $(\epsilon', k\delta + \delta')$-DP for*

$$\epsilon' = \epsilon \sqrt{2k \ln(\frac{1}{\delta'})} + k\epsilon(e^\epsilon - 1) \qquad \forall \delta' > 0$$

Proof. Theorem 3.3 in [DGV10] □

Note that for $\epsilon < 1$, ϵ' is $\mathcal{O}(\sqrt{k}\epsilon)$ by approximating $e^\epsilon \approx 1 + \epsilon$. Theorem 4 also implies hoe to set parameters of each mechanism for k-fold adaptive composition to achieve overall privacy budget of ϵ.

Corollary 5. *If $M : \mathbb{N}^{|X|} \to R^k$ is a k-fold adaptive composition of $(\frac{\epsilon}{\sqrt{8k \ln(\frac{1}{\delta})}}, 0)$-DP mechanisms for $\epsilon < 1$, then M is (ϵ, δ)-DP.*

2 Sparse Vector

Sparse Vector Mechanism allows ϵ to degrade in $\mathcal{O}(\epsilon \ln(k))$, which means that you can answer to exponentially more queries for the same fixed privacy budget, compared to basic composition. The Sparse Vector Mechanism takes a long stream of queries and run the Laplace mechanism on each query while only outputting the noisy answers whose noise is above some noisy (Laplace) threshold.
This is especially useful if the analyst believes that only a small number of the queries will have large (or interesting) answers. For example, the queries can be "What's the correlation between A and B?" where most of the elements are uncorrelated.
Following algorithm 1 is a simplified version called Above Noisy Threshold introduced in [DNR$^+$09]. The Above Noisy Threshold takes a stream of queries and halts after finding the first query with noisy answer above a noisy threshold.

Theorem 6. *Above Noisy Threshold is (ϵ, δ)-DP.*

Proof. [DNR$^+$09] and Theorem 3.23 in the textbook [DR14] (p57) □

Algorithm 1: Above Noisy Threshold $(x, \{f_i\}, T, \epsilon)$

Above Noisy Threshold $(x, \{f_i\}, T, \epsilon)$;
Input : database x, adaptively chosen stream of sensitivity-1 queries $\{f_i\}$, threshold T, privacy parameter ϵ
Output: Stream of answers $\{a_i\} \in \{\perp, \top\}^*$
Let $\hat{T} = T + Lap(\frac{2}{\epsilon})$
for *each query* f_i **do**
  Let $v_i = Lap(\frac{4}{\epsilon})$;
  if $f_i(x) + v_i \geq \hat{T}$ **then**
    output $a_i = \top$;
    Halt ;
  else
    output $a_i = \perp$;
  end
end

The following definition of accuracy and theorem 8 indicate that the mechanism produces an output for all k queries and that its output is approximately correct for all queries with high probability.

Definition 7. *(Accuracy) A mechanism that outputs a stream of answers $\{a_i\} \in \{\perp, \top\}^*$ to a stream of k queries $\{f_i\}$ is (α, β)-accurate with respect to a threshold T if, with probability $\geq 1 - \beta$, the mechanism does not halt before f_k, and*

$$\forall a_i = \top : f_i(x) \geq T - \alpha$$
$$\forall a_i = \perp : f_i(x) \leq T + \alpha$$

Theorem 8. *For any sequence of k sensitivity-1 queries $f_1, \ldots, f_k$ s.t. $|\{i < k : f_i(x) \geq T - \alpha\}| = 0$, then Above Noisy Threshold is (α, β) accurate for*

$$\alpha = \frac{8(\ln(k) + \ln(\frac{2}{\beta})}{\epsilon}$$

Proof. [DNR+09] and Theorem 3.24 in the textbook [DR14] (p59) □

Note that in Theorem 8, the qualifier $|\{i < k : f_i(x) \geq T - \alpha\}| = 0$ requires that the only query close to being above threshold is possibly the last one. Without this condition, the algorithm would be required to halt before the k^{th} query with high probability, so it may bot satisfy the accuracy guarantee.

References

[DGV10] Cynthia Dwork, Rothblum N. Guy, and Salil Vadhan. Boosting and differential privacy. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on. IEEE*, pages 51–60, 2010.

[DNR+09] Cynthia Dwork, Moni Naor, Omer Reingold, Rothblum M. Guy, and Salil Vadhan. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390, 2009.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(34):211–407, 2014.