Exploring Composition Bound in Non-Adaptive Setting

Chan Kang[*]

Harvard University
chankang@college.harvard.edu

August 10, 2016

Abstract

Composition of differentially private algorithms is a central problem in differential privacy since often it is necessary to make multiple queries on a single database or make queries on multiple databases containing a common individual. It is known that the comsposition of differentially private algorithms is again differentially private. Naturally, we would be interested in the parameters of the resulting algorithm in terms of the parameters of the algorithms used in composition. The existing composition theorems have the constraint that the parameters of algorithms being composed be given up front. We refer to this setting – the one where the parameters are given up front – as non-adaptive. On the other hand, one can consider the other setting where the analyst can adjust the parameters based on the previous output. We refer to this setting as adaptive.

As of now, the composition bound in the adaptive setting hasn't been explored as much as in the non-adaptive setting. In fact, before this paper, it was not known whether the non-adaptive composition theorem bound held for the adaptive setting. Here, we describe the experiment we ran to show that the non-adaptive advanced composition bound does not hold for the adaptive case. Lastly, we also discuss some additional experiments we ran to find the scalar that, when multiplied to the original advanced composition theorem bound, makes the bound tight.

I. Introduction

i. Preliminaries

The concept of differential privacy has emerged as a theoretical framework that quantifies the privacy loss of answering a query. The precise definition of an algorithm to be (ϵ, δ)-differentially private is given by the following:

Definition I.1. $[DR10]$ A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|\mathcal{X}|}$ is (ϵ, δ)-differentially private if for all $\mathcal{S} \subseteq \text{Range}\ (\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|\mathcal{X}|}$ such that $\|x - y\|_1 \leq 1$:

$$\Pr\left[\mathcal{M}(x) \in \mathcal{S}\right] \leq \exp(\epsilon) \Pr\left[\mathcal{M}(y) \in \mathcal{S}\right] + \delta$$

In fact, (ϵ, δ)-differential privacy is approximately equivalent to the condition that for all neighboring x, y, the absolute value of the privacy loss is bounded by ϵ with probability at least $1 - \delta$.

A concept closely related to differentially private algorithms is that of privacy loss:

Definition I.2. $[DR10]$

$$\mathcal{L}^{(\xi)}_{\mathcal{M}(x)\|\mathcal{M}(y)} = \ln\left(\frac{\Pr[\mathcal{M}(x)=\xi]}{\Pr[\mathcal{M}(y)=\xi]}\right)$$

the above quantity is referred to as *privacy loss* incurred by observing ξ. It is easy to see that $\mathcal{M}$ is ϵ-differentially private if and only if $L^{(\xi)}_{\mathcal{M}(x)\|\mathcal{M}(y)}$ is bounded by ϵ for all $\xi \in \text{Range}(\mathcal{M})$,

Now we describe two different settings that we will be discussing the composition bounds

[*] Harvard University, anticipated concentration in Mathematics, class of 2017

in. First of the two settings, which we refer to as the *non-adaptive* setting, the parameters of algorithms to be composed are determined in advance. In contrast, in the *adaptive* setting, the adversary can adjust the parameters of later algorithms based on the output of the algorithms that he/she has seen so far. For a more formal, rigorous description of these two settings, see $[RRUV16]$.

Before we continue the discussion of these two settings and composition bounds in each, we state what is often referred to as "basic composition theorem" as well as "advanced composition theorem."

Theorem I.1. (Basic Composition [DMNS06]) The sequence $\epsilon_1, \ldots, \epsilon_k$ and $\delta_1, \ldots, \delta_k$ satisfies (ϵ_g, δ_g)-differential privacy under adaptive composition where

$$\epsilon_g = \sum_{i=1}^k \epsilon_i \text{ and } \delta_g = \sum_{i=1}^k \delta_i$$

Theorem I.2. (Advanced Composition [DRV10]) For any $\hat{\delta} > 0$, the sequence $\epsilon_1, \ldots, \epsilon_k$ and $\delta_1, \ldots, \delta_k$ where $\epsilon = \epsilon_i$ and $\delta = \delta_i$ for all $i \in [k]$ satisfies (ϵ_g, δ_g)-differential privacy under adaptive composition where

$$\epsilon_g = \epsilon(e^\epsilon - 1)k + \epsilon\sqrt{2k\log(1/\hat{\delta})}, \text{ and}$$
$$\delta_g = k\delta + \hat{\delta}$$

This can be generalized to the case where not all the ϵ_i's are the same $[KOV15]$.

Note from the statements of these theorems that the parameters of individual algorithms are all given up front. In other words, these theorems hold in the non-adaptive setting. However, in the adaptive setting, we did not know whether the advanced composition theorem bound held for the adaptive setting. In the remainder of this paper, we describe our experiment, which showed that the advanced composition bound for the non-adaptive setting does not hold for the adaptive setting.

ii. The connection between asymmetric random walks and composition

Consider the map $M : \{0,1\} \to \{0,1\}$ defined by

$$M(b) = \begin{cases} b & \text{with probability } \frac{e^\epsilon}{1+e^\epsilon} \\ b \oplus 1 & \text{with probability } \frac{1}{1+e^\epsilon} \end{cases}$$

It is known that M can be used to used to simulate the output of every ϵ-DP algorithm on adjacent databases.

Lemma I.3 (KOV15)**.** For every ϵ-DP algorithm $\tilde{M}$ and neighboring databases D_0, D_1, there exists a randomized algorithm T such that $T(M(b))$ is identically distributed to $\tilde{M}(D_b)$ for $b = 0$ and $b = 1$.

Hence, it suffices to study the composition of multiple, independent M's – the result can be generalized to composition of any differentially private algorithms.

In our experiment, we compose k independent ϵ-differentially private M, and refer to the resulting algorithm as M'. In other words, M' is a map from $\{0,1\}$ to $\{0,1\}^k$ where each component is an independent M.

Now we consider the random variable $L_{M'}$ that denotes the absolute value of privacy loss of M' over all the possible outcomes of M'. In other words, we are considering

$$\begin{aligned} L_{M'} &= \ln \frac{\Pr[M'(0) = \mathbf{b}]}{\Pr[M'(1) = \mathbf{b}]} \\ &= \ln \frac{\prod_{i=1}^k \Pr[M(0) = b_i]}{\prod_{i=1}^k \Pr[M(1) = b_i]} \\ &= \sum_{i=1}^k \frac{\Pr[M(0) = b_i]}{\Pr[M(1) = b_i]} \\ &= \sum_{i=1}^k L_{M_i} \end{aligned}$$

Note that we used the fact that each coordinate of the output of M' is independent of each other when going from the first line to the second line.

Here, we notice that $L_{M'}$ is essentially the sum of k independent random variables where each one of these k random variables is the random variable L_{M_i} for the privacy loss of the algorithm M. Since

$$L_{M_i} = \begin{cases} \epsilon & \text{with probability } \frac{e^{\epsilon}}{1+e^{\epsilon}} \\ -\epsilon & \text{with probability } \frac{1}{1+e^{\epsilon}} \end{cases}$$

$|L_{M'}|$, which is the absolute value of the sum of k of L_{M_i}, is the distance between the origin and where the random walk ends up after k steps, where the random walk goes up by ϵ with probability $\frac{e^{\epsilon}}{1+e^{\epsilon}}$ and down by ϵ with probability $\frac{1}{1+e^{\epsilon}}$.

II. Methods

i. Non-adaptive Advanced Composition Bound in Adaptive Setting

Here, we reiterate the first result: the advanced composition bound for the non-adaptive setting does *not* hold in the adaptive setting. To obtain this result, we ran n instances of random walks of length k, with some δ value set in advance (we used $\delta = 0.0001$). Let

$$f(\epsilon, \delta, k) = \epsilon(e^{\epsilon} - 1)k + \epsilon\sqrt{2k \log(1/\delta)}$$

Each simulation of random walk consisted of the following: we checked whether the distance from the origin to the current position of the random walk exceeded $f(k, \epsilon, \delta)$. If it ever did, we counted that instance of random walk as "failure." If not, we continued until the random walk made the k^{th} step.

If the non-adaptive advanced composition bound held in adaptive setting, the number of failures n_f divided by n should be approximately δ, since the privacy loss should be more than $f(\epsilon, \delta, k)$ with probability δ.

ii. Finding the right scalar

Once we found that the original advanced composition bound (for the non-adaptive setting) was too tight for the adaptive setting, we started running the same experiments but with the bound scaled up by a constant factor – we already knew from the theoretical work in [RRUV16] that some scaling of the original advanced composition bound (for the non-adaptive setting) holds for the adaptive setting.

III. Results and Discussion

i. Non-adaptive Advanced Composition Bound in Adaptive Setting

The value of δ needed to be sufficiently small, so we gave it a value of 0.0001. To get a sufficiently large number of failures (note that the expected number of failures would be $\delta \cdot n$), we had to make n as large as 500000. Lastly, we set $\epsilon = 0.0001$ and $k = 1000$. The number of failures we got in the first experiment was 230, which was significantly more than the expected number of failures (50) by more than 10 standard deviations: Note that the number of failures follows the binomial distribution with parameters (n, δ). Since the value of n is sufficiently large, we used confidence interval test as if this distribution were Gaussian. The standard deviation is

$$n \cdot \delta \cdot (1 - \delta) = 500000 \cdot 0.0001 \cdot 0.9999 \approx 7.07$$

Hence, we concluded that the advanced composition bound for the non-adaptive setting does not hold for the adaptive setting.

ii. Finding the right scalar

By trial and error, we found that scaling the original bound by 1.08 works reasonably well – i.e. with the bound scaled up by 1.08, the value n_f matches $n\delta$ quite closely, within a standard deviation (in fact, within half the standard deviation in most cases).

Below, we give a table of experimental results.

Scalar	ϵ	δ	n	n_f	k
1.5	0.001	0.001	10000	0	50000
1.1	0.001	0.0001	290000	14	1000
1.05	0.001	0.0001	200000	37	1000
1.05	0.001	0.0001	200000	28	2000
1.08	0.001	0.0001	500000	49	1000
1.08	0.001	0.0001	500000	51	2000
1.08	0.001	0.0001	500000	49	2000
1.08	0.001	0.0001	500000	49	3000
1.08	0.001	0.0001	500000	50	4000
1.08	0.001	0.0001	500000	72	4000
1.08	0.001	0.0001	500000	59	5000

iii. Ideas for Future Research

The bound

$$\epsilon(e^{\epsilon}-1)k+\epsilon\sqrt{2k\log(1/\delta)}$$

given by the advanced composition theorem has two terms – interpreted in the context of random walk, the first term indicates the expected distance away from the origin, and the second term the deviation from the expected position. We reasoned that in both the non-adaptive and the adaptive setting the expected distance away from the origin is the same and hence perhaps the right way to scale the bound is just to scale up the second (deviation) term, rather than both of the terms, as we did in our experiment. When tried with the same scalar 1.08 multiplied just to the second term, the number of failures was more than we wanted it to be (for the given δ).

Scalar	ϵ	δ	n	n_f	k
1.08	0.001	0.0001	500000	34	1000
1.08	0.0001	0.0001	500000	37	1000
1.08	0.001	0.0001	500000	63	2000
1.08	0.001	0.0001	500000	68	3000
1.08	0.001	0.0001	500000	63	4000
1.08	0.001	0.0001	500000	80	5000
1.08	0.0001	0.0001	500000	76	5000
1.08	0.00001	0.0001	500000	74	5000

References

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In TCC ’06, pages 265-284, 2006.

[DKM$^+$06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006*, pages 586-503. Springer, 2006.

[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 51-60, 2010.

[DR10] Cynthia Dwork, Aaron Roth. The Algorithmic Foundations of Differential Privacy. Foundations and Trends in Theoretical Computer Science Vol. 9. Nos. 3-4 (2014)

[RRUV16] Ryan Rogers, Aaron Roth, Jonathan Ullman, Salil Vadhan. Privacy Odometers and Filters: Pay-as-you-Go Composition.

[KOV15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1376-1385, 2015.