Individual Privacy Accounting via a Rényi Filter

Vitaly Feldman
Apple

Tijana Zrnic*
University of California, Berkeley

Abstract

We consider a sequential setting in which a single dataset of individuals is used to perform adaptively-chosen analyses, while ensuring that the differential privacy loss of each participant does not exceed a pre-specified privacy budget. The standard approach to this problem relies on bounding a worst-case estimate of the privacy loss over all individuals and all possible values of their data, for every single analysis. Yet, in many scenarios this approach is overly conservative, especially for "typical" data points which incur little privacy loss by participation in most of the analyses. In this work, we give a method for tighter privacy loss accounting based on the value of a personalized privacy loss estimate for each individual in each analysis. The accounting method relies on a new composition theorem for Rényi differential privacy, which allows adaptively-chosen privacy parameters. We apply our results to the analysis of noisy gradient descent and show how existing algorithms can be generalized to incorporate individual privacy accounting and thus achieve a better privacy-utility tradeoff.

1 Introduction

Understanding how privacy of an individual degrades as the number of analyses using their data grows is of paramount importance in privacy-preserving data analysis. On one hand, this allows individuals to participate in multiple disjoint statistical analyses, all the while knowing that their privacy cannot be compromised by aggregating the resulting reports. On the other hand, this feature is crucial for privacy-preserving algorithm design — instead of having to reason about the privacy properties of a complex algorithm, it allows reasoning about the privacy of the subroutines that make up the final algorithm.

For differential privacy [11], this accounting of privacy losses is typically done using composition theorems. Importantly, given that statistical analyses often rely on the outputs of previous analyses, and that algorithmic subroutines feed into one another, the composition theorems need to be *adaptive*, namely, allow the choice of which algorithm to run next to depend on the outputs of all previous computations. For example, in gradient descent, the computation of the gradient depends on the value of the current iterate, which itself is the output of the previous steps of the algorithm.

Given the central role that adaptive composition theorems play for differentially private data analysis, they have been investigated in numerous works (e.g. [14, 22, 9, 26, 25, 3, 28, 5, 29]). While they differ in some aspects, they also share one limitation. Namely, all of these theorems reason about the worst-case privacy loss for each constituent algorithm in the composition. Here, "worst-case" refers to the worst choice of individual in the dataset and worst choice of value for their data. This pessimistic accounting implies that every algorithm is summarized via a single privacy parameter, shared among all participants in the analysis.

In most scenarios, however, different individuals have different effects on each of the algorithms, as measured by differential privacy. More precisely, the output of an analysis may have little to no dependence on the presence of some individuals. For example, if we wish to report the average income in a neighborhood, removing an individual whose income is close to the average has virtually no impact on the final report after noise addition. Similarly, when training a machine learning model via gradient

*Work done while at Apple.

descent, the norm of the gradient of the loss function defined by a data point is often much smaller than the maximum norm (typically determined by a clipping operation). As a result, in many cases no single individual is likely to have the worst-case effect on all the steps of the analysis. This means that accounting based on existing composition theorems may be unnecessarily conservative.

In this work, we present a tighter analysis of privacy loss composition by computing the associated divergences at an individual level. In particular, to achieve a pre-specified privacy budget, we keep track of a personalized estimate of the divergence for each individual in the analyzed dataset, and ensure that the respective estimate for all individuals is maintained under the budget throughout the composition. Additionally, even when there is no a priori budget on privacy, we design algorithms that provide a valid running upper bound on the privacy loss for each individual in the dataset.

The rest of the paper is organized as follows. In the remainder of this section, we give an overview of our main results and discuss related work. In the next section, we introduce the preliminaries necessary to state our results, and give an illustration of the benefits of individual privacy accounting in the setting of non-adaptive composition. In Section 3, we prove our main adaptive composition theorem for Rényi differential privacy. We build off this result in Section 4 and Section 5, where we develop Rényi privacy filters and odometers — objects for budgeting and tracking privacy loss — and apply them to individual privacy accounting. In Section 6 we present an application of our theory to differentially private optimization, as well as some experimental results.

1.1 Overview of main results

It is straightforward to measure the worst-case effect of a specific data point on a given analysis in terms of any of the divergences used to define differential privacy. One can simply replace the supremum over all datasets in the standard definition of (removal) differential privacy with the supremum over datasets that include that specific data point (see Def. 2.4). Indeed, such a definition was given in the work of Ebadi et al. [17] and a related definition is given by Wang [30]. However, a meaningful application of adaptive composition with such a definition immediately runs into the following technical challenge. Standard adaptive composition theorems require that the privacy parameter of each step be fixed in advance. For individual privacy parameters, this approach requires using the worst-case value of the individual privacy loss over all the possible analyses at a given step. Individual privacy parameters tend to be much more sensitive to the analysis being performed than worst-case privacy losses, and thus using the worst-case value over all analyses is likely to negate the benefits of using individual privacy losses in the first place.

Thus the main technical challenge in analyzing composition of individual privacy losses is that they are themselves random variables that depend on the outputs of all the previous computations. More specifically, if we denote by $a_1,\dots,a_{t-1}$ the output of the first $t-1$ adaptively composed algorithms $\mathcal{A}_1,\dots,\mathcal{A}_{t-1}$, then the individual privacy loss of any point incurred by applying algorithm $\mathcal{A}_t$ is a function of $a_1,\dots,a_{t-1}$.Therefore, to tackle the problem of composing individual privacy losses we need to understand composition with *adaptively-chosen* privacy parameters in general. We refer to this kind of composition as *fully adaptive*.

The setting of fully adaptive privacy composition is rather subtle and even defining privacy in terms of the adaptively-chosen privacy parameters requires some care. This setting was first studied by Rogers et al. [28] who introduced the notion of a *privacy filter*. Informally, a privacy filter is a stopping time rule that halts a computation based on the adaptive sequence of privacy parameters and ensures that a pre-specified privacy budget is not exceeded. Rogers et al. define a filter for approximate differential privacy that asymptotically behaves like the advanced composition theorem [14], but is substantially more involved and loses a constant factor. Moreover, several of the tighter analyses of Gaussian noise addition require composition to be done in Rényi differential privacy [1, 25]. Converting them to (ε,δ)-differential privacy would incur an additional $\sqrt{\log(1/\delta)}$ factor in the final bound.

Our main result can be seen as a privacy filter for Rényi differential privacy (RDP) whose stopping rule exactly matches the rate of standard RDP composition [25].

Theorem 1.1. *Fix any $B \geqslant 0, \alpha \geqslant 1$. Suppose that $\mathcal{A}_t$ is (α, ρ_t)-Rényi differentially private, where ρ_t is an arbitrary function of $a_1, \ldots, a_{t-1}$. If $\sum_{t=1}^{k} \rho_t \leqslant B$ holds almost surely, then the adaptive composition of $\mathcal{A}_1, \ldots, \mathcal{A}_k$ is (α, B)-Rényi differentially private.*

This theorem implies that stopping the algorithm based on the sum of privacy parameters so far is valid even under fully adaptive composition. Note that, when all privacy parameters are fixed, Theorem 1.1 recovers the usual composition result for RDP [25]. Our RDP filter immediately implies a simple filter for approximate differential privacy that is as tight as any version of the advanced composition theorem obtained via concentrated differential privacy [3] (see Thm. 4.5). These Rényi-divergence-based composition analyses are known to improve upon the rate of Dwork et al. [14] and, in particular, improve on the results in [28].

Rogers et al. [28] also define a *privacy odometer*, which provides an upper bound on the running privacy loss, and does not require a pre-specified budget. We show that simply adding up Rényi privacy parameters does not make for a valid odometer for RDP, however by applying a discretization argument, we construct an approximate odometer using our Rényi privacy filter.

We instantiate our general results for fully adaptive composition in the setting of individual privacy accounting. This allows us to define an *individual privacy filter*, which, given a fixed privacy budget, adaptively drops points from the analysis once their *personalized* privacy loss estimate exceeds the budget. Therefore, instead of keeping track of a single running privacy loss estimate for all individuals, we track a less conservative, personalized estimate for each individual in the dataset. Individual privacy filtering allows for better, adaptive utilization of data points for a given budget, which we also demonstrate in experiments.

1.2 Related work

The main motivation behind our work is obtaining tighter privacy accounting methods through, broadly speaking, "personalized" accounting of privacy losses. Existing literature in differential privacy discusses several related notions [20, 17, 30, 4], although typically with an incomparable objective. Ghosh and Roth [20] discuss individual privacy in the context of selling privacy at auction and their definition does not depend on the value of the data point but only on its index in the dataset. Cummings and Durfee [4] rely on a similar privacy definition, investigate an associated definition of individual sensitivity, and demonstrate a general way to preprocess an arbitrary function of a dataset into a function that has the desired bounds on individual sensitivities.

Ebadi et al. [17] introduce personalized differential privacy in the context of private database queries and describe a system which drops points when their personalized privacy loss exceeds a budget. In their system personalized privacy losses result from record selection operations applied to the database. While this type of accounting is similar to ours in spirit, their work only considers basic and non-adaptive composition. The work of Wang [30] considers the privacy loss of a specific data point relative to a fixed dataset. It provides techniques for evaluating this "per-instance" privacy loss for several statistical problems. Wang [30] also briefly discusses adaptive composition of per-instance differential privacy as a straightforward generalization of the usual advanced composition theorem [14], but the per-instance privacy parameters are assumed to be *fixed*. As discussed above, having fixed per-instance privacy parameters, while allowing adaptive composition, is likely to negate some of the benefits of personalized privacy estimates. The work of Ligett et al. [24] tightens individuals' personalized privacy loss by taking into account subsets of analyses in which an individual does not participate. Within the studies in which an individual participates, however, they consider the usual worst-case privacy loss, rather than an individual one. In addition, the analyses in which the user participates are determined in a data-independent way.

Our work can be seen as related to data-dependent approaches to analysis of privacy-preserving algorithms such as smooth sensitivity [27], the propose-test-release framework [7], and *ex-post* privacy guarantees [31]. The focus of our work is complementary in that we aim to capture the dependence of the output on the value of each individual's data point as opposed to the "easiness" of the entire dataset. Our approach also requires composition to exploit the gains from individual privacy loss accounting.

Finally, adaptive composition of differentially private algorithms is one of the key tools for establishing statistical validity of an adaptively-chosen sequence of statistical analyses [16, 15, 2]. In this context, Feldman and Steinke [19] show that the individual KL-divergence losses (or RDP losses for $\alpha = 1$) compose adaptively and can be exploited for deriving tighter generalization results. However, their results still require that the average of individual KL-divergences be upper bounded by a fixed worst-case value and the analysis appears to be limited to the $\alpha = 1$ case.

2 Preliminaries

We will denote by $S = (X_1, \dots, X_n)$ the analyzed dataset, and by $S^{-i} \stackrel{\text{def}}{=} (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n)$ the analyzed dataset after removing point X_i. We will generally focus on algorithms that can take as input a dataset or arbitrary size. If, instead, the algorithm requires an input of fixed size, one can obtain the same results for algorithms that replace X_i with an arbitrary element $X^\star$ fixed in advance (for example 0).

We start by reviewing some preliminaries on differential privacy.

Definition 2.1 ([11, 10])**.** A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private (DP) if for all datasets $S = (X_1, \dots, X_n)$,

$$\mathbb{P}\mathrm{r}\left[\mathcal{A}(S) \in E\right] \leqslant e^{\varepsilon}\, \mathbb{P}\mathrm{r}\left[\mathcal{A}(S^{-i}) \in E\right] + \delta, \text{ and } \mathbb{P}\mathrm{r}\left[\mathcal{A}(S^{-i}) \in E\right] \leqslant e^{\varepsilon}\, \mathbb{P}\mathrm{r}\left[\mathcal{A}(S) \in E\right] + \delta,$$

for all $i \in [n]$ and all measurable sets E.

The technical portion of our analysis will make use of Rényi differential privacy (RDP), a relaxation of DP based on Rényi divergences which often leads to tighter privacy bounds than analyzing DP directly. Formally, the Rényi divergence of order $\alpha \in (1, \infty)$ between two measures μ and ν such that $\mu \ll \nu$ is defined as:

$$D_\alpha(\mu \| \nu) = \frac{1}{\alpha - 1} \log \int \left(\frac{d\mu}{d\nu}\right)^{\alpha} d\nu.$$

The Rényi divergence of order $\alpha = 1$ is defined by continuity, and recovers the Kullback-Leibler (KL) divergence. Relying on a standard abuse of notation, for two random variables X and Y, we will use $D_\alpha(X\|Y)$ to denote the Rényi divergence between their distributions.

Similarly as in Definition 2.1, to argue that an algorithm is Rényi differentially private, we will need to bound both $D_\alpha(\mathcal{A}(S)\|\mathcal{A}(S^{-i}))$ and $D_\alpha(\mathcal{A}(S^{-i})\|\mathcal{A}(S))$, for all $i \in [n]$. For this reason, we introduce shorthand notation for the maximum of the two directions of Rényi divergence:

$$D_\alpha^{\leftrightarrow}(X\|Y) \stackrel{\text{def}}{=} \max\{D_\alpha(X\|Y), D_\alpha(Y\|X)\}.$$

Definition 2.2 ([25])**.** A randomized algorithm $\mathcal{A}$ is (α, ρ)-Rényi differentially private (RDP) if for all datasets $S = (X_1, \dots, X_n)$,

$$D_\alpha^{\leftrightarrow}\left(\mathcal{A}(S) \| \mathcal{A}(S^{-i})\right) \leqslant \rho,$$

for all $i \in [n]$.

Rényi differential privacy implies differential privacy; therefore, although our guarantees will be stated in terms of RDP, the conversion to DP is immediate.

Fact 2.3 ([25])**.** *If algorithm $\mathcal{A}$ is (α, ρ)-RDP, then it is also $\left(\rho + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right)$-DP, for any $\delta \in (0, 1)$.*

One of the successes of differential privacy (and RDP as well) lies in its adaptive composition property. In Algorithm 1 we define adaptive composition, which is at the center of our analysis.

Algorithm 1 Adaptive composition $\mathcal{A}^{(k)}$

input: dataset $S \in \mathcal{X}^n$, sequence of algorithms $\mathcal{A}_t, t = 1, 2, \dots, k$
for $t = 1, \dots, k$ **do**
  Compute $a_t = \mathcal{A}_t(a_1, \dots, a_{t-1}, S)$
end
Return $\mathcal{A}^{(k)}(S) = (a_1, \dots, a_k)$

If $\mathcal{A}_t(a_1, \dots, a_{t-1}, \cdot)$ is (α, ρ_t)-RDP for *all* values of $a_1, \dots, a_{t-1}$, then the standard adaptive composition theorem for RDP says that $\mathcal{A}^{(k)}$ is $(\alpha, \sum_{t=1}^{k} \rho_t)$-RDP [25]. Note that, by definition, the parameters $(\rho_1, \dots, \rho_k)$ are independent of the specific $a_1, \dots, a_{t-1}$ obtained in the adaptive computation. In other words, they are fixed in advance.

2.1 Individual privacy losses

Our individual accounting relies on measuring the maximum possible effect of an individual data point on a dataset statistic in terms of Rényi divergence. This measure is equivalent to an RDP version of personalized differential privacy [17]. For convenience we will refer to it as individual Rényi differential privacy, or individual RDP for short. We note, however, that, by itself, a bound on this divergence does not imply any formal privacy guarantee for the individual. In particular, an individual RDP parameter itself depends on the sensitive value of the data point.

Definition 2.4 (Individual RDP)**.** Fix $n \in \mathbb{N}$ and data point X. We say that a randomized algorithm $\mathcal{A}$ satisfies (α, ρ)-*individual Rényi differential privacy for* X if for all datasets $S \supseteq X$ such that $|S| \leqslant n$, it holds that

$$D_\alpha^{\leftrightarrow}(\mathcal{A}(S) \| \mathcal{A}(S \setminus X)) \leqslant \rho.$$

Therefore, to satisfy the standard definition of RDP, an algorithm needs to satisfy individual RDP for all data points X.

Our main focus will be on individual privacy losses as introduced in Definition 2.4, however some of our results also hold under a weaker notion of individual privacy loss, which measures the effect of a data point on the output of a statistical analysis, relative to a *fixed* dataset. This notion is an RDP version of per-instance differential privacy [30].

Definition 2.5 (Individual RDP (per-instance))**.** Fix a dataset $S = (X_1, \dots, X_n)$. We say that a randomized algorithm $\mathcal{A}$ satisfies (α, ρ)-*individual Rényi differential privacy for* (S, X_i) if it holds that

$$D_\alpha^{\leftrightarrow}\left(\mathcal{A}(S) \| \mathcal{A}(S^{-i})\right) \leqslant \rho.$$

We will note which results hold under Definition 2.5, in addition to being valid under Definition 2.4.

Before we turn to analyzing composition, we give a simple example of individual RDP computation.

Example 2.6 (Linear queries)**.** Let $S = (X_1, \dots, X_n) \in \mathcal{X}^n$. Suppose that $\mathcal{A}$ is a d-dimensional linear query with Gaussian noise addition, $\mathcal{A}(S) = \sum_{j \in [n]} \phi(X_j) + \xi$, for some $\phi \colon \mathcal{X} \to \mathbb{R}^d$ and $\xi \sim N(0, \sigma^2 \mathbb{I}_d)$. Then, $\mathcal{A}$ satisfies

$$\left(\alpha, D_\alpha^{\leftrightarrow}\left(N\left(\sum_{j \in [n]} \phi(X_j), \sigma^2 \mathbb{I}_d\right) \middle\| N\left(\sum_{j \in [n], j \neq i} \phi(X_j), \sigma^2 \mathbb{I}_d\right)\right)\right) = \left(\alpha, \frac{\alpha \|\phi(X_i)\|_2^2}{2\sigma^2}\right)$$

individual RDP for X_i. Note that in this case individual RDP (Definition 2.4) and per-instance RDP (Definition 2.5) have the same value.

The analysis above extends to arbitrary Lipschitz functions.

Example 2.7 (Lipschitz analyses)**.** Suppose that $g : (\mathbb{R}^d)^n \to \mathbb{R}^{d'}$ is L_i-Lipschitz in coordinate i. For $\phi \colon \mathcal{X} \to \mathbb{R}^d$, let $\mathcal{A}(S) = g(\phi(X_1), \dots, \phi(X_n)) + \xi$, $\xi \sim N(0, \sigma^2 \mathbb{I}_{d'})$. Assume that for some $X^\star$, $\phi(X^\star)$ is the origin. Then, by using $X^\star$ to replace a removed element (namely $S^{-i} = (X_1, \dots, X_{i-1}, X^\star, X_{i+1}, \dots, X_n)$), we get that $\mathcal{A}$ satisfies $\left(\alpha, \frac{\alpha L_i^2 \|\phi(X_i)\|_2^2}{2\sigma^2}\right)$ individual RDP for X_i.

In the previous examples we focus on Gaussian noise for simplicity. Similar computations can be carried out for other randomization mechanisms.

The main high-level idea behind our approach in tightening the composition analysis is to keep track of the individual privacy losses for all samples $X_i \in S$ at every round of privacy composition. Before we delve into the analysis of adaptive composition through the lens of individual privacy, we first consider a simpler setting, where all computations are non-adaptively-chosen.

2.2 Warm-up: tighter non-adaptive composition via individual privacy

To illustrate the gains of using individual privacy, we first consider non-adaptive composition, where the computation at time t has no dependence on the past reports; that is, $\mathcal{A}_t(a_1, \ldots, a_{t-1}, S) \equiv \mathcal{A}_t(S)$.

For simplicity, suppose that we sequentially receive bounded queries $q_1, q_2, \ldots$, where $q_t : \mathcal{X} \to [0,1]$ for all $t \in \mathbb{N}$, and we want to report their sum over a dataset $S = (X_1, \ldots, X_n) \in \mathcal{X}^n$ subject to the constraint that the output needs to be (α, B)-RDP, for some pre-specified budget B. A prototypical mechanism for answering such queries is the Gaussian mechanism, which reports $a_t = \sum_{i=1}^n q_t(X_i) + \xi_t$, where $\xi_t \sim N(0, \sigma^2)$. This mechanism satisfies $(\alpha, \rho_t^{(i)})$ individual RDP for X_i, where

$$\rho_t^{(i)} = D_\alpha^{\leftrightarrow}\left(N\left(\sum_{j\in[n]} q_t(X_j), \sigma^2\right), N\left(\sum_{j\in[n], j\neq i} q_t(X_j), \sigma^2\right)\right) = \frac{\alpha q_t(X_i)^2}{2\sigma^2}.$$

In contrast, the worst-case differential privacy parameters are $(\alpha, \rho_t) = \left(\alpha, \frac{\alpha}{2\sigma^2}\right)$.

Given the constraint of the overall report being (α, B)-Rényi differentially private, classical analyses — which only consider ρ_t — would suggest answering at most $k_0 = \lfloor \frac{2B\sigma^2}{\alpha} \rfloor$ queries via the Gaussian mechanism, in order to ensure privacy.

By using individual privacy, we can ensure the same privacy guarantees while achieving a potentially higher utility in terms of the number of answered queries. Indeed, the number of queries that can be answered accurately is determined by the queries evaluated on the *analyzed* data.

Proposition 2.8. *For all $t \in \mathbb{N}$, let $S_t = \left(X_i \in S : \sum_{j=1}^t q_j(X_i)^2 \leqslant \frac{2B\sigma^2}{\alpha}\right)$, and let $a_t = \sum_{X_i \in S_t} q_t(X_i) + \xi_t$, where $\xi_t \sim N(0, \sigma^2)$. Then, the sequence of reports $a_1, a_2, \ldots$ is (α, B)-Rényi differentially private. Moreover, for all t such that $\sum_{j=1}^t q_j(X_i)^2 \leqslant \frac{2B\sigma^2}{\alpha} \forall i$, it holds that $|a_t - \sum_{X_i\in S} q_t(X_i)| \leqslant \sqrt{2\log(1/\delta)}\sigma$ with probability at least $1-\delta$.*

Here, S_t denotes the set of "active" points at time t. Note that, due to $\rho_t^{(i)} \leqslant \rho_t$, *all* points are active for at least the first k_0 computations, as prescribed by the usual worst-case analysis. Therefore, individual privacy provides a more fine-grained way of quantifying privacy loss by taking into account the value of the point whose loss we aim to measure. Naturally, after a certain number of reports we expect few points to remain active, and hence one needs to decide on a stopping criterion for the reports. We discuss stopping criteria in more detail later on.

To give an example application of Proposition 2.8, in the context of continual monitoring [13, 18], a platform might collect one real-valued indicator per user per day, and wish to make decisions based off the daily averages of these indicators across users. Here, X_i would be a single user, and $q_t(X_i) = X_t^{(i)}$ would be the corresponding user's indicator on day t. Assuming Gaussian noise addition, Proposition 2.8 suggests that the platform can utilize the data of each user in a privacy-preserving manner as long as the squared ℓ_2-norm of their indicators is not too large. This implication of Proposition 2.8 is especially important in applications where the collected values are naturally sparse; for example, if the collected values are binary indicators $X_t^{(i)} \in \{0,1\}$ of a change of state. While a classical approach to privacy would allow a total number of reports proportional to the privacy budget, an individual privacy approach allows making a number of reports proportional to the number of positive reports, that is $\sum_{j=1}^t X_j^{(i)}$, for all individuals i.

Proposition 2.8 is easy to prove using standard techniques, by considering the ℓ_2-sensitivity of the entire output of the algorithm. This result also holds for more general analyses (with $\rho_t^{(i)}$ used in place of

$\frac{\alpha q_j(X_i)^2}{2\sigma^2}$ in the computation of the "active" dataset S_t) and the proof follows directly from known properties of composition in RDP. In our main result we argue that the same result holds when composition is *adaptive*.

3 Fully adaptive composition for Rényi differential privacy

Our main technical contribution is a new adaptive composition theorem for Rényi differential privacy, which bounds the overall privacy loss in terms of the individual privacy losses of all data points. As argued earlier, the main challenge in understanding how individual privacy parameters compose is the fact that these parameters are random, rather than fixed. In what follows, we first state a general version of our main theorem, which bounds the privacy loss in adaptive composition in terms of a bound on the sequence of possibly random privacy parameters. Then, we instantiate this result in the context of individual privacy.

We set up some notation within the context of adaptive composition (Algorithm 1). We denote by $a^{(t)} \stackrel{\text{def}}{=} (a_1, \ldots, a_t)$ the sequence of the first t reports, and by $\mathcal{A}^{(t)}(\cdot) \stackrel{\text{def}}{=} (\mathcal{A}_1(\cdot), \mathcal{A}_2(\mathcal{A}_1(\cdot), \cdot), \ldots, \mathcal{A}_t(\mathcal{A}_1(\cdot), \ldots, \cdot))$ the composed algorithm which produces $a^{(t)}$. For two datasets S and S', parameter $\alpha \geqslant 1$ and *fixed* $a^{(t)}$, we let[1]

$$\mathrm{Loss}^{(t)}(a^{(t)}; S, S', \alpha) \stackrel{\text{def}}{=} \left(\frac{\mathbb{P}\mathrm{r}\left[\mathcal{A}^{(t)}(S) = a^{(t)}\right]}{\mathbb{P}\mathrm{r}\left[\mathcal{A}^{(t)}(S') = a^{(t)}\right]} \right)^{\alpha}.$$

Similarly, for fixed $a^{(t)}$ we also define

$$\mathrm{Loss}_t(a^{(t)}; S, S', \alpha) \stackrel{\text{def}}{=} \left(\frac{\mathbb{P}\mathrm{r}\left[\mathcal{A}_t(a_1, \ldots, a_{t-1}, S) = a_t\right]}{\mathbb{P}\mathrm{r}\left[\mathcal{A}_t(a_1, \ldots, a_{t-1}, S') = a_t\right]} \right)^{\alpha}.$$

Roughly speaking, $\mathrm{Loss}^{(t)}$ denotes the total privacy loss incurred by the first t rounds of adaptive composition, while Loss_t denotes the loss incurred in round t (which, due to adaptivity, depends on the outcomes of the first $t-1$ rounds).

Generally, we will be interested in $\mathrm{Loss}_t(a^{(t)}; S, S', \alpha)$ and $\mathrm{Loss}^{(t)}(a^{(t)}; S, S', \alpha)$ when $a^{(t)}$ is output by adaptive composition; in such cases, these two quantities are random.

Note that, since

$$\begin{aligned}\mathbb{P}\mathrm{r}\left[\mathcal{A}^{(t)}(S) = a^{(t)}\right] &= \mathbb{P}\mathrm{r}\left[\mathcal{A}^{(t-1)}(S) = a^{(t-1)}\right] \mathbb{P}\mathrm{r}\left[\mathcal{A}_t(\mathcal{A}_1(S), \ldots, \mathcal{A}_{t-1}(\mathcal{A}_1(S), \ldots), S) = a_t \mid \mathcal{A}^{(t-1)}(S) = a^{(t-1)}\right] \\ &= \mathbb{P}\mathrm{r}\left[\mathcal{A}^{(t-1)}(S) = a^{(t-1)}\right] \mathbb{P}\mathrm{r}\left[\mathcal{A}_t(a_1, a_2, \ldots, a_{t-1}, S) = a_t\right],\end{aligned}$$

we have $\mathrm{Loss}^{(t)}(a^{(t)}; S, S', \alpha) = \mathrm{Loss}^{(t-1)}(a^{(t-1)}; S, S', \alpha) \cdot \mathrm{Loss}_t(a^{(t)}; S, S', \alpha)$.

We let ρ_t denote the RDP parameter of order α of $\mathcal{A}_t$, conditional on the past reports. For the sake of generality and simplicity of exposition, we introduce an abstract space $\mathcal{S}$ over pairs of datasets and let

$$\rho_t = \frac{1}{\alpha - 1} \log \sup_{(S,S') \in \mathcal{S}} \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{(t)}(S')} \left[\mathrm{Loss}_t(a^{(t)}; S, S', \alpha) \,\middle|\, a^{(t-1)} \right]. \tag{1}$$

In the context of individual privacy (Definition 2.4), we will instantiate $\mathcal{S}$ to be the space of all dataset pairs where either dataset is obtained by deleting X_i from the other. For the per-instance notion (Definition 2.5), we will set $\mathcal{S} = \left\{(S, S^{-i}), (S^{-i}, S)\right\}$. In the context of usual RDP, $\mathcal{S}$ will be the space of all pairs of datasets that differ in the presence of one element.

The classical composition theorem for Rényi differential privacy — while allowing $\mathcal{A}_t$ to depend on the previous reports — constrains $\mathcal{A}_t$ to be (α, ρ_t)-RDP for some *fixed* ρ_t. Here, we make no such constraints on ρ_t; hence, ρ_t will in general be a random variable, due to the randomness in $a_1, \ldots, a_{t-1}$.

[1] All algorithms we will be considering in this paper, if not discrete, induce a density with respect to the Lebesgue measure. For such instances, replacing expressions such as $\mathbb{P}\mathrm{r}\left[\mathcal{A}^{(t)}(S) = a\right]$ with the density of $\mathcal{A}^{(t)}(S)$ at a gives the analysis in the continuous case.

Theorem 3.1 states that, as long as $\sum_{t=1}^{k} \rho_t$ is maintained under a fixed budget, the output of adaptive composition preserves privacy.

Theorem 3.1. *Fix any $B \geqslant 0, \alpha \geqslant 1$, and a set of pairs of datasets $\mathcal{S}$. If $\sum_{t=1}^{k} \rho_t \leqslant B$ holds almost surely, then the adaptive composition $\mathcal{A}^{(k)}$ satisfies*

$$D_\alpha\left(\mathcal{A}^{(k)}(S) \| \mathcal{A}^{(k)}(S')\right) \leqslant B,$$

for all $(S, S') \in \mathcal{S}$.

Proof. Fix any $(S, S') \in \mathcal{S}$. In what follows, we take $a^{(t)} = (a_1, \ldots, a_t)$ to be distributed as the random output of adaptive composition applied to S', that is $a^{(t)} \sim \mathcal{A}^{(t)}(S')$. Consequently, $\mathrm{Loss}^{(t)}(a^{(t)}; S, S', \alpha)$ and $\mathrm{Loss}_t(a^{(t)}; S, S', \alpha)$ are also random.

Let $M_t \stackrel{\text{def}}{=} \mathrm{Loss}^{(t)}(a^{(t)}; S, S', \alpha) e^{-(\alpha-1)\sum_{j=1}^{t} \rho_j}$, and let $M_0 = 1$. Consider the filtration $\mathcal{F}_t = \sigma(a^{(t)})$. We prove that M_t is a supermartingale with respect to $\mathcal{F}_t$; that is, we show $\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] \leqslant M_{t-1}$. This follows since:

$$\begin{aligned}
\mathbb{E}[M_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\mathrm{Loss}^{(t)}(a^{(t)}; S, S', \alpha)\, e^{-(\alpha-1)\sum_{j=1}^{t} \rho_j} \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathrm{Loss}^{(t-1)}(a^{(t-1)}; S, S', \alpha)\, \mathrm{Loss}_t(a^{(t)}; S, S', \alpha)\, e^{-(\alpha-1)\sum_{j=1}^{t} \rho_j} \,\middle|\, \mathcal{F}_{t-1}\right] \\
&= \mathrm{Loss}^{(t-1)}(a^{(t-1)}; S, S', \alpha)\, e^{-(\alpha-1)\sum_{j=1}^{t} \rho_j}\, \mathbb{E}\left[\mathrm{Loss}_t(a^{(t)}; S, S', \alpha) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\leqslant \mathrm{Loss}^{(t-1)}(a^{(t-1)}; S, S', \alpha)\, e^{-(\alpha-1)\sum_{j=1}^{t} \rho_j} e^{(\alpha-1)\rho_t} \\
&= \mathrm{Loss}^{(t-1)}(a^{(t-1)}; S, S', \alpha)\, e^{-(\alpha-1)\sum_{j=1}^{t-1} \rho_j} \\
&= M_{t-1},
\end{aligned}$$

where the third equality uses the fact that $(\rho_j)_{j=1}^{t} \in \mathcal{F}_{t-1}$, and the inequality applies the definition of ρ_t. Therefore, by applying iterated expectations, we can conclude

$$\mathbb{E}[M_k] = \mathop{\mathbb{E}}_{a^{(k)} \sim \mathcal{A}^{(k)}(S')}\left[\mathrm{Loss}^{(k)}\left(a^{(k)}; S, S', \alpha\right) e^{-(\alpha-1)\sum_{j=1}^{k} \rho_j}\right] \leqslant \mathbb{E}[M_0] = 1.$$

Since $\sum_{j=1}^{k} \rho_j \leqslant B$ by assumption, this inequality implies

$$\mathop{\mathbb{E}}_{a^{(k)} \sim \mathcal{A}^{(k)}(S')}\left[\mathrm{Loss}^{(k)}\left(a^{(k)}; S, S', \alpha\right)\right] \leqslant e^{(\alpha-1)B}.$$

After normalizing, we get

$$D_\alpha\left(\mathcal{A}^{(k)}(S), \mathcal{A}^{(k)}(S')\right) = \frac{1}{\alpha-1} \log \mathop{\mathbb{E}}_{a^{(k)} \sim \mathcal{A}^{(k)}(S')}\left[\mathrm{Loss}^{(k)}\left(a^{(k)}; S, S', \alpha\right)\right] \leqslant B.$$

Since the choice of (S, S') was arbitrary, we can conclude $\sup_{(S,S') \in \mathcal{S}} D_\alpha\left(\mathcal{A}^{(k)}(S), \mathcal{A}^{(k)}(S')\right) \leqslant B$, as desired. ∎

We now instantiate Theorem 3.1 in the context of individual privacy.

We use $\rho_t^{(i)}$ to denote the individual privacy parameter of the t-th adaptively composed algorithm $\mathcal{A}_t$ with respect to X_i, conditional on the past reports. More formally, for fixed $\alpha \geqslant 1$ and for any data point $X_i \in S$ we let:

$$\rho_t^{(i)} \stackrel{\text{def}}{=} \frac{1}{\alpha-1} \log \sup_{S \supseteq X_i : |S| \leqslant n} \max\left\{ \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{(t)}(S^{-i})}\left[\mathrm{Loss}_t(a^{(t)}; S, S^{-i}, \alpha) \,\middle|\, a^{(t-1)}\right], \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{(t)}(S)}\left[\mathrm{Loss}_t(a^{(t)}; S^{-i}, S, \alpha) \,\middle|\, a^{(t-1)}\right] \right\}. \tag{2}$$

Since $\rho_t^{(i)}$ is an instance of the general definition (1) obtained by specifying $\mathcal{S}$, a direct corollary of Theorem 3.1 is as follows.

Corollary 3.2. *Fix any $B \geqslant 0$. If for any input dataset $S = (X_1, \dots, X_n)$, $\sum_{t=1}^k \rho_t^{(i)} \leqslant B$ holds almost surely for all individuals $i \in [n]$, then the adaptive composition $\mathcal{A}^{(k)}$ is (α, B)-Rényi differentially private.*

Proof. By Theorem 3.1, $\sum_{t=1}^k \rho_t^{(i)} \leqslant B$ implies that

$$D_\alpha^{\leftrightarrow}\left(\mathcal{A}^{(k)}(S) \| \mathcal{A}^{(k)}(S^{-i})\right) \leqslant B.$$

Since this holds for all $i \in [n]$, we conclude that $\mathcal{A}^{(k)}$ is (α, B)-Rényi differentially private. ∎

Notice that $\sum_{t=1}^k \rho_t^{(i)} \leqslant B$ is a *data-specific* requirement, while classical composition results consider *all* hypothetical datasets. In Subsection 4.1 we will show how Corollary 3.2 can be operationalized.

It is worth mentioning that Corollary 3.2 also holds under the per-instance notion of individual privacy (Definition 2.5). This result is obtained by simply taking $\mathcal{S} = \{(S, S^{-i}), (S^{-i}, S)\}$ in the proof, where S is the analyzed dataset. However, our main application of Corollary 3.2 — individual privacy filtering, stated in the following section — requires individual privacy loss accounting according to Definition 2.4.

4 Rényi privacy filter

Rogers et al. [28] define the notion of a *privacy filter*, an object that takes as input adaptively-chosen DP parameters $\varepsilon_1, \delta_1, \dots, \varepsilon_t, \delta_t$, as well as a global differential privacy budget ε_g, δ_g, and outputs CONT if the overall report after t rounds of adaptive composition with the corresponding privacy parameters is guaranteed to satisfy $(\varepsilon_g, \delta_g)$-DP. Otherwise, it outputs HALT.

By an immediate extension of Theorem 3.1, we show that a valid RDP version of a privacy filter simply adds up privacy parameters, as in the usual composition where all privacy parameters are fixed up front. By existing conversions of RDP guarantees to DP guarantees, we recover a filter for DP as well, thus improving upon the rate obtained by Rogers et al.

As in equation (1), we let ρ_t denote the possibly random RDP parameter of order α of $\mathcal{A}_t$, conditional on the past reports. We again assume an implicit space $\mathcal{S}$ over pairs of datasets, which we instantiate in different ways depending on the privacy accounting method.

Algorithm 2 Adaptive composition with Rényi privacy filtering

input: dataset $S \in \mathcal{X}^n$, maximum number of rounds $N \in \mathbb{N}$, sequence of algorithms $\mathcal{A}_k, k = 1, 2, \dots, N$

Initialize $k = 0$

Compute $\rho_1 = \frac{1}{\alpha-1} \log \ \sup_{(S,S') \in \mathcal{S}} \ \mathbb{E}_{a^{(1)} \sim \mathcal{A}^{(1)}(S')} \left[\mathrm{Loss}_1(a^{(1)}; S, S', \alpha)\right]$

while $\mathcal{F}_{\alpha,B}(\rho_1, \dots, \rho_{k+1}) = \mathrm{CONT}$ *and* $k < N$ **do**

  $k \leftarrow k + 1$

  Compute $a_k = \mathcal{A}_k(a_1, \dots, a_{k-1}, S)$

  Compute $\rho_{k+1} = \frac{1}{\alpha-1} \log \ \sup_{(S,S') \in \mathcal{S}} \ \mathbb{E}_{a^{(k+1)} \sim \mathcal{A}^{(k+1)}(S')} \left[\mathrm{Loss}_{k+1}(a^{(k+1)}; S, S', \alpha) \;\middle|\; a^{(k)}\right]$

end

Return $\mathcal{A}^{(k)}(S) = (a_1, \dots, a_k)$

Let S_∞ denote the set of all positive, real-valued finite sequences.

Definition 4.1 (RDP filter)**.** Fix a parameter $\alpha \geqslant 1$, and privacy budget B. We say that $\mathcal{F}_{\alpha,B} : S_\infty \to \{\mathrm{CONT}, \mathrm{HALT}\}$ is a valid Rényi privacy filter, or RDP filter for short, if for any sequence of algorithms $(\mathcal{A}_k)_{k=1}^N$, Algorithm 2 satisfies

$$D_\alpha^{\leftrightarrow}\left(\mathcal{A}^{(k)}(S) \| \mathcal{A}^{(k)}(S^{-i})\right) \leqslant B,$$

for all datasets $S = (X_1, \dots, X_n)$ and $i \in [n]$.

We make a remark about the use of privacy filters. The analyst might choose an algorithm at time t that exceeds the privacy budget, which will trigger the filter $\mathcal{F}_{\alpha,B}$ to halt. However, the analyst can then decide to change the computation at time t retroactively and query the filter again, which then might allow continuation. This way, one can ensure a sequence of N computations with formal privacy guarantees, for any target number of rounds N. In the following subsection, we present an application of RDP filters to individual privacy loss accounting which relies on this reasoning.

Theorem 4.2. *Let*

$$\mathcal{F}_{\alpha,B}(\rho_1,\dots,\rho_k) = \begin{cases} \text{CONT}, & \text{if } \sum_{t=1}^{k} \rho_t \leq B, \\ \text{HALT}, & \text{if } \sum_{t=1}^{k} \rho_t > B. \end{cases}$$

Then, $\mathcal{F}_{\alpha,B}$ is a valid Rényi privacy filter.

Proof. The only difference between Theorem 3.1 and this theorem is that a privacy filter halts at a random round, meaning the length of the output is random rather than fixed. Therefore, in this proof we formalize the fact that Theorem 3.1 is valid even under adaptive stopping.

Fix any $(S,S') \in \mathcal{S}$. By the argument in Theorem 3.1, $M_t \stackrel{\text{def}}{=} \text{Loss}^{(t)}(a^{(t)};S,S',\alpha)e^{-(\alpha-1)\sum_{j=1}^{t}\rho_j}$, where $a^{(t)} \sim \mathcal{A}^{(t)}(S')$, is a supermartingale with respect to $\mathcal{F}_t = \sigma(a^{(t)})$.

Let T be the time when Algorithm 2 halts; that is,

$$T = \min\{t : \mathcal{F}_{\alpha,B}(\rho_1,\dots,\rho_{t+1}) = \text{HALT}\} \wedge N.$$

Note that T is a stopping time with respect to $\mathcal{F}_t$, that is $\{T = t\} \in \mathcal{F}_t$, due to the fact that $\rho_{t+1} \in \mathcal{F}_t$. Since T is almost surely bounded by construction, we can apply the optional stopping theorem for supermartingales to get

$$\mathbb{E}[M_T] = \mathop{\mathbb{E}}_{a^{(T)}\sim\mathcal{A}^{(T)}(S')}\left[\text{Loss}^{(T)}\left(a^{(T)};S,S',\alpha\right)e^{-(\alpha-1)\sum_{j=1}^{T}\rho_j}\right] \leq \mathbb{E}[M_0] = 1.$$

By definition of the RDP filter, we know that $\sum_{j=1}^{T}\rho_j \leq B$ almost surely; otherwise the filter would have halted earlier. Thus, we can conclude

$$\mathop{\mathbb{E}}_{a^{(T)}\sim\mathcal{A}^{(T)}(S')}\left[\text{Loss}^{(T)}\left(a^{(T)};S,S',\alpha\right)e^{-(\alpha-1)B}\right] \leq 1.$$

After rearranging and normalizing, this implies

$$D_\alpha\left(\mathcal{A}^{(T)}(S),\mathcal{A}^{(T)}(S')\right) = \frac{1}{\alpha-1}\log \mathop{\mathbb{E}}_{a^{(T)}\sim\mathcal{A}^{(T)}(S')}\left[\text{Loss}^{(T)}\left(a^{(T)};S,S',\alpha\right)\right] \leq B.$$

Therefore, $\mathcal{F}_{\alpha,B}$ is a valid RDP filter. ∎

We remark that Rogers et al. define a privacy filter somewhat more generally, by treating the analyst as an adversary who is allowed to pick "bad" neighboring datasets at every step (see Algorithm 2 in [28]). Theorem 4.2 holds under this setting as well, however we opted for a simpler presentation.

Just like Corollary 3.2 applies Theorem 3.1 in the individual privacy setting, we can apply Rényi privacy filters to individual RDP parameters, in which case the filter indicates whether the privacy loss of a specific individual is potentially violated.

4.1 Individual privacy filter

Now we design an *individual privacy filter*, which monitors individual privacy loss estimates across all individuals and all computations, and ensures that the privacy of all individuals is preserved. The filter guarantees privacy by adaptively dropping data points once their cumulative individual privacy loss estimate is about to cross a pre-specified budget. More specifically, at every step of adaptive composition

t, it determines an active set of points $S_t \subseteq S$ based on cumulative estimated individual losses, and applies $\mathcal{A}_t$ only to S_t.

Algorithm 3 Adaptive composition with individual privacy filtering

input: dataset $S \in \mathcal{X}^n$, sequence of algorithms $\mathcal{A}_t$, $t = 1, 2, \ldots, k$
for $t = 1, \ldots, k$ **do**
  For all $X_i \in S$, compute $\rho_t^{(i)}$ (as in eq. (2))
  Determine active set $S_t = \left(X_i \;:\; \mathcal{F}_{\alpha,B}(\rho_1^{(i)}, \ldots, \rho_t^{(i)}) = \mathrm{CONT}\right)$
  For all $X_i \in S$, set $\rho_t^{(i)} \leftarrow \rho_t^{(i)} \mathbf{1}\{X_i \in S_t\}$
  Compute $a_t = \mathcal{A}_t(a_1, \ldots, a_{t-1}, S_t)$
end
Return $(a_1, \ldots, a_k)$

Here, $\mathcal{F}_{\alpha,B}$ is the Rényi privacy filter from Theorem 4.2. Given its validity, one can observe that Algorithm 3 preserves Rényi differential privacy.

Theorem 4.3. *Adaptive composition with individual privacy filtering (Algorithm 3) satisfies (α, B)-Rényi differential privacy.*

Proof. Denote by $\mathcal{A}_t^{\text{filt}}$ the subroutine given by the t-th step of the individual filtering algorithm; that is, $a_t = \mathcal{A}_t^{\text{filt}}(a_1, \ldots, a_{t-1}, S)$. Note that $\mathcal{A}_t^{\text{filt}}$ is *not* equal to $\mathcal{A}_t$. By analogy with the notation $\mathcal{A}^{(t)}$, we also let $\mathcal{A}^{\text{filt}(t)}(\cdot) \stackrel{\text{def}}{=} (\mathcal{A}_1^{\text{filt}}(\cdot), \mathcal{A}_2^{\text{filt}}(\mathcal{A}_1^{\text{filt}}(\cdot), \cdot), \ldots, \mathcal{A}_t^{\text{filt}}(\mathcal{A}_1^{\text{filt}}(\cdot), \ldots, \cdot))$.

We argue that the privacy loss of point X_i in round t, conditional on the past reports, is upper bounded by $\rho_t^{(i)}$ (after $\rho_t^{(i)}$ has been updated):

$$\frac{1}{\alpha-1} \log \max\left\{ \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{\text{filt}(t)}(S^{-i})}\left[\mathrm{Loss}_t^{\text{filt}}(a^{(t)}; S, S^{-i}, \alpha) \;\middle|\; a^{(t-1)}\right], \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{\text{filt}(t)}(S)}\left[\mathrm{Loss}_t^{\text{filt}}(a^{(t)}; S^{-i}, S, \alpha) \;\middle|\; a^{(t-1)}\right] \right\} \leqslant \rho_t^{(i)}, \tag{3}$$

where $\mathrm{Loss}_t^{\text{filt}}(a^{(t)}; S, S', \alpha) = \left(\frac{\mathbb{P}\mathrm{r}\left[\mathcal{A}_t^{\text{filt}}(a_1, \ldots, a_{t-1}, S) = a_t\right]}{\mathbb{P}\mathrm{r}\left[\mathcal{A}_t^{\text{filt}}(a_1, \ldots, a_{t-1}, S') = a_t\right]}\right)^\alpha$.

To do so, we reason about the active set of points at time t when the input to adaptive composition is S, and when the input is S^{-i}. Denote by S_t the active set given input S, and by S_t^{-i} the active set given input S^{-i}. Observe that, conditional on $a_1, \ldots, a_{t-1}$, $S_t^{-i} = S_t \setminus X_i$. This follows because the sequence $(\rho_j^{(i)})_{j=1}^t$ is measurable with respect to $a_1, \ldots, a_{t-1}$, and whether point X_i is active at time t is in turn determined based only on $(\rho_j^{(i)})_{j=1}^t$. In particular, whether any given point is active does not depend on the rest of the input dataset (S or S^{-i}), given $a_1, \ldots, a_{t-1}$. Therefore, if $X_i \notin S_t$, then X_i loses no privacy in round t, because $\mathcal{A}_t^{\text{filt}}(a_1, \ldots, a_{t-1}, S) \stackrel{d}{=} \mathcal{A}_t^{\text{filt}}(a_1, \ldots, a_{t-1}, S^{-i})$, conditional on $a_1, \ldots, a_{t-1}$. On the other hand, if $X_i \in S_t$, then its privacy loss can be bounded as

$$\begin{aligned}
&\frac{1}{\alpha-1} \log \max\left\{ \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{\text{filt}(t)}(S^{-i})}\left[\mathrm{Loss}_t^{\text{filt}}(a^{(t)}; S, S^{-i}, \alpha) \;\middle|\; a^{(t-1)}\right], \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{\text{filt}(t)}(S)}\left[\mathrm{Loss}_t^{\text{filt}}(a^{(t)}; S^{-i}, S, \alpha) \;\middle|\; a^{(t-1)}\right] \right\} \\
\leqslant\; &\frac{1}{\alpha-1} \log \max\left\{ \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{\text{filt}(t)}(S^{-i})}\left[\mathrm{Loss}_t(a^{(t)}; S_t, S_t^{-i}, \alpha) \;\middle|\; a^{(t-1)}\right], \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{\text{filt}(t)}(S)}\left[\mathrm{Loss}_t^{\text{filt}}(a^{(t)}; S_t^{-i}, S_t, \alpha) \;\middle|\; a^{(t-1)}\right] \right\} \\
\leqslant\; &\frac{1}{\alpha-1} \log \sup_{S \supseteq X_i} \max\left\{ \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{(t)}(S^{-i})}\left[\mathrm{Loss}_t(a^{(t)}; S, S^{-i}, \alpha) \;\middle|\; a^{(t-1)}\right], \mathop{\mathbb{E}}_{a^{(t)} \sim \mathcal{A}^{(t)}(S)}\left[\mathrm{Loss}_t(a^{(t)}; S^{-i}, S, \alpha) \;\middle|\; a^{(t-1)}\right] \right\} \\
\leqslant\; &\rho_t^{(i)}.
\end{aligned}$$

With this, we have showed that $\rho_t^{(i)}$ is a valid estimate of the privacy loss of X_i, for all $i \in [n]$.

Now we argue that, at the end of every round t (after $\rho_t^{(i)}$ has been updated), $\mathcal{F}_{\alpha,B}(\rho_1^{(i)},\ldots,\rho_t^{(i)}) = \text{CONT}$ for all $i \in [n]$. This follows by induction. For $t = 1$, this is clearly true because $\mathcal{F}_{\alpha,B}(0) = \text{CONT}$. Now assume it is true at time $t-1$. Then, at time t, the filter clearly continues for all $X_i \in S_t$ simply by definition of S_t. If $X_i \notin S_t$, then $\mathcal{F}_{\alpha,B}(\rho_1^{(i)},\ldots,\rho_t^{(i)}) = \mathcal{F}_{\alpha,B}(\rho_1^{(i)},\ldots,\rho_{t-1}^{(i)},0) = \mathcal{F}_{\alpha,B}(\rho_1^{(i)},\ldots,\rho_{t-1}^{(i)}) = \text{CONT}$. Therefore, we conclude that at the end of every round $t \in [k]$ and all $i \in [n]$, the filter would output CONT. By the validity of $\mathcal{F}_{\alpha,B}$, we know that $\mathcal{F}_{\alpha,B}(\rho_1^{(i)},\ldots,\rho_t^{(i)}) = \text{CONT}$ implies

$$D_\alpha^{\leftrightarrow}\left(\mathcal{A}^{(t)}(S)\|\mathcal{A}^{(t)}(S^{-i})\right) \leqslant B,$$

and since this holds for all S and all $i \in [n]$, we conclude that Algorithm 3 is (α, B)-RDP. ■

We can now justify the use of a supremum over all datasets that include X in Definition 2.4 and, in particular, why the per-instance notion in Definition 2.5 does not suffice. By the current design of Algorithm 3, $X_i \notin S_t$ implies no privacy loss for point X_i. This is true because, conditional on $a_1,\ldots,a_{t-1}$, X_i being inactive ensures that S_t would be the same regardless of whether the input to Algorithm 3 is S or S^{-i}. Consequently, the output at time t would be insensitive to the value of X_i. Under the more fine-grained definition of individual privacy, *even if* $X_i \notin S_t$, its privacy could still leak at round t. The reason is that, under two different inputs S and S^{-i}, the running privacy loss estimates for all points are different, and hence the active set S_t in the two hypothetical scenarios could be different as well. This fact, in turn, implies two different distributions over reports a_t. In short, dropping X_i from the analysis *does not* prevent its further privacy leakage if accounting is done according to Definition 2.5.

In Section 6, we apply the individual privacy filter to differentially private optimization via gradient descent, and demonstrate how this object ensures utilization of data points as long as their *realized* gradients have low norm.

4.2 (ε, δ)-differential privacy filter via Rényi filter

By connections between Rényi differential privacy and approximate differential privacy [3, 25], we can translate our Rényi privacy filter into a filter for approximate differential privacy.

We define a valid DP filter analogously to a valid RDP filter, the difference being that it takes as input DP, rather than RDP parameters, and that it is parameterized by a global DP budget $\varepsilon_g \geqslant 0, \delta_g \in (0,1)$. We denote by ε_t the possibly adaptive differential privacy parameter of $\mathcal{A}_t$:

$$\varepsilon_t = \sup_{(S,S')\in\mathcal{S}} \sup_E \frac{\mathbb{P}\mathrm{r}\left[\mathcal{A}_t(a_1,\ldots,a_{t-1},S) \in E \mid a^{(t-1)}\right]}{\mathbb{P}\mathrm{r}\left[\mathcal{A}_t(a_1,\ldots,a_{t-1},S') \in E \mid a^{(t-1)}\right]}.$$

We focus on advanced composition of *pure* differentially private algorithms $\mathcal{A}_t$. As shown in [28], a DP filter for approximately differentially private algorithms $\mathcal{A}_t$ can be obtained by an immediate extension of a filter for pure DP algorithms.

Algorithm 4 Adaptive composition with differential privacy filtering

input: dataset $S \in \mathcal{X}^n$, maximum number of rounds $N \in \mathbb{N}$, sequence of algorithms $\mathcal{A}_k, k = 1,2,\ldots,N$
Initialize $k = 0$
while $\mathcal{G}_{\varepsilon_g,\delta_g}(\varepsilon_1,\ldots,\varepsilon_{k+1}) = \text{CONT}$ *and* $k < N$ **do**
  $k \leftarrow k+1$
  Compute $a_k = \mathcal{A}_k(a_1,\ldots,a_{k-1},S)$
end
Return $\mathcal{A}^{(k)}(S) = (a_1,\ldots,a_k)$

As before, S_∞ denotes the set of all positive, real-valued finite sequences.

Definition 4.4 (DP filter)**.** Fix $\varepsilon_g \geqslant 0, \delta_g \in (0,1)$. We say that $\mathcal{G}_{\varepsilon_g,\delta_g} : S_\infty \to \{\text{CONT}, \text{HALT}\}$ is a valid differential privacy filter, or DP filter for short, if for any sequence of algorithms $(\mathcal{A}_k)_{k=1}^N$, Algorithm 4 satisfies

$$\mathbb{P}\mathrm{r}\left[\mathcal{A}^{(k)}(S) \in E\right] \leqslant e^{\varepsilon_g}\,\mathbb{P}\mathrm{r}\left[\mathcal{A}^{(k)}(S^{-i}) \in E\right] + \delta_g, \quad \mathbb{P}\mathrm{r}\left[\mathcal{A}^{(k)}(S^{-i}) \in E\right] \leqslant e^{\varepsilon_g}\,\mathbb{P}\mathrm{r}\left[\mathcal{A}^{(k)}(S) \in E\right] + \delta_g,$$

for all datasets $S = (X_1, \ldots, X_n)$, $i \in [n]$, and measurable events E.

We note that Rogers et al. [28] define a differential privacy filter somewhat differently — for example, in their definition a filter admits a sequence of privacy parameters of fixed length — however Definition 4.4 is essentially equivalent to theirs.

By invoking standard conversions between DP and Rényi-divergence-based privacy notions, our analysis implies a simple stopping condition for a DP filter, in terms of any zero-concentrated differential privacy (zCDP) level which ensures $(\varepsilon_g, \delta_g)$-DP. For clarity, we give one particularly simple such translation from zCDP to DP, however one could in principle invoke more sophisticated analyses such as those of Bun and Steinke [3]. In general, we can reproduce any rate for advanced composition of DP that utilizes Rényi-divergence-based privacy definitions, in the setting of fully adaptive composition.

Theorem 4.5. *Let $B^\star$ be the largest $B > 0$ such that B-zero-concentrated differential privacy (zCDP) implies $(\varepsilon_g, \delta_g)$-differential privacy. Let*

$$\mathcal{G}_{\varepsilon_g,\delta_g}(\varepsilon_1, \ldots, \varepsilon_k) = \begin{cases} \text{CONT}, & \text{if } \frac{1}{2}\sum_{t=1}^k \varepsilon_t^2 \leqslant B^\star, \\ \text{HALT}, & \text{if } \frac{1}{2}\sum_{t=1}^k \varepsilon_t^2 > B^\star. \end{cases}$$

Then, $\mathcal{G}_{\varepsilon_g,\delta_g}$ is a valid DP filter. For example,

$$\mathcal{G}_{\varepsilon_g,\delta_g}(\varepsilon_1, \ldots, \varepsilon_k) = \begin{cases} \text{CONT}, & \text{if } \frac{1}{2}\sum_{t=1}^k \varepsilon_t^2 \leqslant \left(-\sqrt{\log(1/\delta_g)} + \sqrt{\log(1/\delta_g) + \varepsilon_g}\right)^2, \\ \text{HALT}, & \text{if } \frac{1}{2}\sum_{t=1}^k \varepsilon_t^2 > \left(-\sqrt{\log(1/\delta_g)} + \sqrt{\log(1/\delta_g) + \varepsilon_g}\right)^2 \end{cases}$$

is a valid DP filter.

Proof. By conversions between DP and zCDP [3], we know that ε_t-DP implies $\frac{1}{2}\varepsilon_t^2$-zCDP, that is $(\alpha, \frac{1}{2}\varepsilon_t^2\alpha)$-RDP, for all $\alpha \geqslant 1$. Thus, a Rényi filter with parameters $(\alpha, \alpha B^\star)$ would stop once $\frac{1}{2}\sum_{t=1}^k \varepsilon_t^2 > B^\star$. Since this condition is independent of α, the output of adaptive composition with this stopping condition satisfies $(\alpha, B^\star)$-RDP for all $\alpha \geqslant 1$. This guarantee is equivalent to $B^\star$-zCDP, and by assumption this implies $(\varepsilon_g, \delta_g)$-DP as well.

By Fact 2.3, $B^\star$-zCDP implies $\left(min_\alpha \alpha B^\star + \frac{\log(1/\delta_g)}{\alpha-1}, \delta_g\right)$-DP. Optimizing over α and solving for $B^\star$ such that $min_\alpha \alpha B^\star + \frac{\log(1/\delta_g)}{\alpha-1} = \varepsilon_g$ yields $B^\star = \left(-\sqrt{\log(1/\delta_g)} + \sqrt{\log(1/\delta_g) + \varepsilon_g}\right)^2$. ∎

If the privacy parameters are fixed up front and $\varepsilon_t \equiv \varepsilon$, simplifying the stopping criterion of the above DP filter implies that adaptive composition of k ε-differentially private algorithm satisfies

$$\left(\frac{1}{2}k\varepsilon^2 + \sqrt{2k\log(1/\delta)}\varepsilon, \delta\right)\text{-differential privacy,}$$

for all $\delta > 0$. This tightens the rate of Rogers et al. [28]. While reading off the exact rate is difficult, since the smallest ε_g for which the filter continues cannot be easily expressed, their rate is greater than $\left(\frac{1}{2}k\varepsilon(e^\varepsilon - 1) + \sqrt{2k\log(1/\delta)}\varepsilon, \delta\right)$ by a significant constant factor. Further improvements on the rate are possible via a more intricate conversion between RDP (or, rather, zCDP) and DP, as presented in [3].

5 Rényi privacy odometer

A privacy filter is meant to estimate whether the realized privacy loss surpassed a fixed threshold. As such, it is meant to shape the course of adaptive composition by limiting the privacy loss. A *privacy odometer* [28], on the other hand, is a more sophisticated object meant to give a valid upper bound on the privacy loss incurred thus far, typically without constraining the analyses.

First we discuss privacy odometers in the general context of random privacy parameters, and then we apply this general theory toward designing an individual privacy odometer.

Formalizing what it means to have a valid odometer is a delicate task. Perhaps the most natural requirement for an odometer O_t would be for the following condition to hold true almost surely, for all rounds $t \in \mathbb{N}$, datasets $S \in \mathcal{X}^n$, and its elements $i \in [n]$:

$$\frac{1}{\alpha-1}\max\left\{\log \mathop{\mathbb{E}}_{a^{(t)}\sim\mathcal{A}^{(t)}(S^{-i})}\left[\mathrm{Loss}^{(t)}(a^{(t)};S,S^{-i},\alpha)\,\middle|\,O_t\right], \log \mathop{\mathbb{E}}_{a^{(t)}\sim\mathcal{A}^{(t)}(S)}\left[\mathrm{Loss}^{(t)}(a^{(t)};S^{-i},S,\alpha)\,\middle|\,O_t\right]\right\} \leqslant O_t. \quad (4)$$

Given the validity of a privacy filter which simply adds up RDP parameters, it seems reasonable to have an odometer which adds up RDP parameters as well. However, if we set $O_t = \sum_{j=1}^{t} \rho_j$, it is not hard to see that condition (4) does not hold.

Example 5.1. Suppose we run two algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$, which take as input a single bit $S \in \{0,1\}$. Let

$$\rho_1 \stackrel{\text{def}}{=} \frac{1}{\alpha-1}\log \mathop{\mathbb{E}}_{a_1\sim\mathcal{A}^{(1)}(1)}[\mathrm{Loss}_1(a_1;0,1,\alpha)], \ \ \rho_2 \stackrel{\text{def}}{=} \frac{1}{\alpha-1}\log \mathop{\mathbb{E}}_{a^{(2)}\sim\mathcal{A}^{(2)}(1)}\left[\mathrm{Loss}_2((a_1,a_2);0,1,\alpha)\,\middle|\,a_1\right].$$

Suppose that $\mathcal{A}_1$ outputs non-negative reals, and that we pick $\mathcal{A}_2$ such that $\rho_2 = a_1$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{a^{(2)}\sim\mathcal{A}^{(2)}(1)}\left[\mathrm{Loss}^{(2)}(a^{(2)};0,1,\alpha)\,\middle|\,\rho_1+\rho_2\right] &= \mathop{\mathbb{E}}_{a^{(2)}\sim\mathcal{A}^{(2)}(1)}\left[\mathrm{Loss}_1(a_1;0,1,\alpha)\cdot\mathrm{Loss}_2((a_1,a_2);0,1,\alpha)\,\middle|\,\rho_2\right] \\
&= \mathop{\mathbb{E}}_{a^{(2)}\sim\mathcal{A}^{(2)}(1)}\left[\mathrm{Loss}_1(a_1;0,1,\alpha)\cdot\mathrm{Loss}_2((a_1,a_2);0,1,\alpha)\,\middle|\,a_1\right] \\
&= \mathrm{Loss}_1(a_1;0,1,\alpha)\mathop{\mathbb{E}}_{a^{(2)}\sim\mathcal{A}^{(2)}(1)}\left[\mathrm{Loss}_2((a_1,a_2);0,1,\alpha)\,\middle|\,a_1\right] \\
&= \mathrm{Loss}_1(a_1;0,1,\alpha)e^{(\alpha-1)\rho_2}.
\end{aligned}$$

Since $\mathbb{E}_{a_1\sim\mathcal{A}^{(1)}(1)}[\mathrm{Loss}_1(a_1;0,1,\alpha)] = e^{(\alpha-1)\rho_1}$, the right-hand is guaranteed to be at most $e^{(\alpha-1)(\rho_1+\rho_2)}$ almost surely only if $\mathrm{Loss}_1(a_1;0,1,\alpha) \equiv \mathbb{E}\,\mathrm{Loss}_1(a_1;0,1,\alpha)$, and this happens only when $\rho_1 = 0$.

5.1 Approximate privacy odometer via privacy filtering

The previous example illustrates how constructing a valid RDP odometer might be non-trivial. However, we make an observation that a valid Rényi privacy filter can be utilized to design an approximate odometer as follows.

We let ρ_t denote the RDP parameter of order α of $\mathcal{A}_t$, conditional on the past reports — as per equation (1).

Algorithm 5 Tracking privacy loss via Rényi privacy odometer

input: dataset $S \in \mathcal{X}^n$, discretization error $\Delta > 0$, sequence of algorithms $\mathcal{A}_t, t = 1, 2, \ldots, k$

Initialize odometer $O_1 = \Delta$

Set $T_{\text{restart}} = 1$

for $t = 1, 2, \ldots, k$ **do**

  Compute $a_t = \mathcal{A}_t(a_1, \ldots, a_{t-1}, S)$

  if $\mathcal{F}_{\alpha,\Delta}(\rho_{T_{\text{restart}}}, \ldots, \rho_t) = \text{HALT}$ **then**

    Augment odometer $O_t \leftarrow O_{t-1} + \Delta$

    Update restart time $T_{\text{restart}} \leftarrow t$

  else

    $O_t \leftarrow O_{t-1}$

  end

end

In words, every time an RDP filter with privacy budget Δ halts, we restart a new filter and augment the odometer by Δ. Here, $\Delta > 0$ is the discretization error of the odometer.

An important question here is how one should go about choosing Δ. If Δ is large, then the odometer is very coarse and inaccurate. On the other end, if Δ is small, the filter might halt very often, and whenever a filter halts we effectively make the upper bound on the odometer a bit looser. Roughly speaking, if we restart at time t we lose a factor of $\Delta - \sum_{j=T_{\text{restart}}}^{t-1} \rho_j$, where T_{restart} is the last restart time before t.

We now state the guarantees of our Rényi privacy odometer. For all $j \in \mathbb{N}$, let T_j denote the j-th time a filter restarts in Algorithm 5. More formally, we can define the sequence $\{T_j\}_j$ recursively as

$$T_j = \min\left\{t > T_{j-1} : \mathcal{F}_{\alpha,\Delta}(\rho_{T_{j-1}}, \ldots, \rho_t) = \text{HALT}\right\}, \text{ where } T_0 = 0.$$

Proposition 5.2. *Suppose that $\rho_j \leqslant \Delta$ almost surely, for all $j \in \mathbb{N}$, and let t be any time such that $T_{k-1} \leqslant t < T_k$. Then, the odometer O_t in Algorithm 5 upper bounds the privacy loss at time t:*

$$\sup_{(S,S')\in\mathcal{S}} D_\alpha\left(\mathcal{A}^{(t)}(S) \| \mathcal{A}^{(t)}(S')\right) \leqslant k\Delta = O_t, \quad \forall i \in [n].$$

Proof. Since the output $(a_1, \ldots, a_t)$ is a post-processing of $(a_1, \ldots, a_{T_k - 1})$, it suffices to prove that the latter output satisfies $(\alpha, k\Delta)$-RDP. The algorithm $\mathcal{A}^{(T_k - 1)}$ can be written as an adaptive composition of k algorithms, each of which outputs $(a_{T_{j-1}}, \ldots, a_{T_j - 1})$, $j \in \{1, \ldots, k\}$. Therefore, by the standard adaptive composition theorem for RDP, it suffices to argue that each of these k algorithms is RDP, conditional on the outputs of the previous algorithms. Since $\mathcal{F}_{\alpha,\Delta}$ is a valid Rényi privacy filter by Theorem 4.2, each of these k algorithms is indeed (α, Δ)-RDP, which completes the proof. ■

5.2 Individual privacy odometer

In the context of individual privacy, Proposition 5.2 allows designing a personalized privacy odometer for all analyzed data points. Here, we track $O_t^{(i)}$ for all points $X_i \in S$. The update is analogous to that of Algorithm 5, the difference being that a separate privacy filter is applied to the individual privacy parameters for all points separately. Naturally, each data point has its own random times of filter exceedances, $\{T_j^{(i)}\}_j$. Formally, we define the sequence $\{T_j^{(i)}\}_j$ recursively as

$$T_j^{(i)} = \min\left\{t > T_{j-1}^{(i)} : \mathcal{F}_{\alpha,\Delta}\left(\rho_{T_{j-1}^{(i)}}^{(i)}, \ldots, \rho_t^{(i)}\right) = \text{HALT}\right\}, \text{ where } T_0^{(i)} = 0.$$

Here, $\rho_t^{(i)}$ are individual privacy parameters, measured according to equation (2).

Algorithm 6 Individual privacy loss tracking via Rényi privacy odometer

input: dataset $S \in \mathcal{X}^n$, discretization error $\Delta > 0$, sequence of algorithms $\mathcal{A}_t, t = 1, 2, \ldots, k$

For all $i \in [n]$, initialize odometer $O_1^{(i)} = \Delta$ and set $T_{\text{restart}}^{(i)} = 1$
for $t = 1, 2, \ldots, k$ **do**
  Compute $a_t = \mathcal{A}_t(a_1, \ldots, a_{t-1}, S)$
  for $i = 1, \ldots, n$ **do**
    if $\mathcal{F}_{\alpha,\Delta}(\rho^{(i)}_{T^{(i)}_{\text{restart}}}, \ldots, \rho_t^{(i)}) = \text{HALT}$ **then**
      Augment odometer $O_t^{(i)} \leftarrow O_{t-1}^{(i)} + \Delta$
      Update restart time $T_{\text{restart}}^{(i)} \leftarrow t$
    else
      $O_t^{(i)} \leftarrow O_{t-1}^{(i)}$
    end
  end
end

It is worth pointing out that odometer values $O_t^{(i)}$ are *sensitive*, as they depend on the value of the data point X_i. Importantly, they can be disclosed to the respective user without violating the other users' privacy; $O_t^{(i)}$ depends on X_i, but it does not depend on the other data points (other than through $a_1, \ldots, a_{t-1}$, which are reported in a privacy-preserving manner).

Below we state an immediate corollary of Proposition 5.2.

Corollary 5.3. *Suppose that $\rho_j^{(i)} \leqslant \Delta$ almost surely, for all $j \in \mathbb{N}$ and $i \in [n]$. For fixed $i \in [n]$, let t be any time such that $T_{k-1}^{(i)} \leqslant t < T_k^{(i)}$. Then, the odometer $O_t^{(i)}$ in Algorithm 6 upper bounds the individual privacy loss of point X_i at time t:*

$$D_\alpha^{\leftrightarrow}\left(\mathcal{A}^{(t)}(S) \| \mathcal{A}^{(t)}(S^{-i})\right) \leqslant k\Delta = O_t^{(i)}.$$

It is worth pointing out that the same odometer validity guarantee would hold if accounting was done according to the per-instance notion of individual privacy (Definition 2.5). In that case, however, odometer values $O_t^{(i)}$ would depend on the whole dataset S, and not just X_i. Consequently, reporting the odometers to users — without violating other users' privacy — would require greater care.

6 Private gradient descent with individual privacy accounting

In this section, we discuss an application of the individual privacy filter from Section 4 to differentially private optimization. Within this application, we also demonstrate the use of Rényi privacy odometers from Section 5.

A popular approach to differentially private model training via gradient descent is to clip the norm of individual gradients at every time step and add Gaussian noise to the clipped gradients [1]. Existing privacy analyses compute the overall privacy spent up to a given round by using a uniform upper bound on the gradient norms, determined by the clipping value. Using the individual privacy filter from Section 4, we develop a less conservative version of private gradient descent, one which takes into account the *realized* norms of the gradients, rather than just their upper bound.

There are various natural ways one could incorporate individual privacy accounting into the standard private gradient descent (GD) algorithm [1]. To facilitate the comparison, we present a particularly simple one. As in private gradient descent, at every step we clip all computed gradients and add Gaussian noise. However, after the round at which private gradient descent would halt, we additionally look at the "leftover" privacy budget for all points, and utilize them until their budget runs out. The leftover budget for each point is essentially equivalent to the difference between the worst-case sum of squared ℓ_2-norms of the gradients (determined by the clipping value) and the sum of squared ℓ_2-norms of the *realized* gradients. Below we provide the details of this application and also contrast it with the standard private gradient descent algorithm.

Algorithm 7 Private gradient descent

input: dataset $(X_1,\dots,X_n)$, loss function $\ell(\theta;X_i)$, learning rate $(\eta_t)_{t=1}^{\infty}$, noise scale $\sigma>0$, clip value $C>0$, number of steps $k\in\mathbb{N}$

Initialize θ_1 arbitrarily

for $t=1,2,\dots,k$ **do**

  Compute gradients $g_t(X_i)\leftarrow\nabla_\theta\ell(\theta_t;X_i),\forall i$

  Clip $\bar{g}_t(X_i)\leftarrow g_t(X_i)\cdot\min\left(1,\frac{C}{\|g_t(X_i)\|_2}\right),\forall i$

  Add noise $\widetilde{g}_t\leftarrow\frac{1}{n}\sum_{i=1}^n(\bar{g}_t(X_i)+N(0,\sigma^2C^2\mathbb{I}))$

  Take gradient step $\theta_{t+1}\leftarrow\theta_t-\eta_t\widetilde{g}_t$

end

Return θ_{k+1}

Algorithm 8 Private gradient descent with filtering

input: dataset $(X_1,\dots,X_n)$, loss function $\ell(\theta;X_i)$, learning rate $(\eta_t)_{t=1}^{\infty}$, noise scale $\sigma>0$, clip value $C>0$, number of steps $k_{\max}\in\mathbb{N}$, squared norm budget $B_{\text{norm}}>0$

Initialize θ_1 arbitrarily

for $t=1,2,\dots,k_{\max}$ **do**

  Compute gradients $g_t(X_i)\leftarrow\nabla_\theta\ell(\theta_t;X_i),\forall i$

  Clip $\bar{g}_t(X_i)\leftarrow$

$$g_t(X_i)\cdot\min\left(1,\frac{\min\left(C,\sqrt{B_{\text{norm}}-\sum_{j=1}^{t-1}\|\bar{g}_j(X_i)\|_2^2}\right)}{\|g_t(X_i)\|_2}\right),\forall i$$

  Add noise $\widetilde{g}_t\leftarrow\frac{1}{n}\sum_{i=1}^n(\bar{g}_t(X_i)+N(0,\sigma^2C^2\mathbb{I}))$

  Take gradient step $\theta_{t+1}\leftarrow\theta_t-\eta_t\widetilde{g}_t$

end

Return θ_{k+1}

In Algorithm 7, all gradients get clipped to have norm at most C at every time step. In Algorithm 8, the gradient for point X_i gets clipped to have norm at most $\min(C,B_{\text{norm}}-\sum_{j=1}^{t-1}\|\bar{g}_j(X_i)\|_2^2)$. This means that, at least for the first $\lfloor B_{\text{norm}}/C^2\rfloor$ rounds, all gradients get clipped to have norm at most C. After round $\lfloor B_{\text{norm}}/C^2\rfloor$, points adaptively get filtered out once the accumulated squared norm of their (clipped) gradients reaches B_{norm}. Therefore, we observe that for $B_{\text{norm}}=kC^2$ and $k_{\max}=k$, Algorithm 8 recovers Algorithm 7.

The standard privacy guarantees of Algorithm 7 are given as follows.

Proposition 6.1. *Private gradient descent (Algorithm 7) satisfies* $\left(\alpha,\frac{\alpha k}{2\sigma^2}\right)$*-RDP, for all* $\alpha\geqslant1$.

We prove the privacy guarantees of Algorithm 8 as a corollary of our individual privacy filter.

Proposition 6.2. *Private gradient descent with filtering (Algorithm 8) satisfies* $\left(\alpha,\frac{\alpha B_{norm}}{2\sigma^2C^2}\right)$*-RDP, for all* $\alpha\geqslant1$.

Proof. By properties of the Gaussian mechanism, the individual RDP parameters of order α are $\rho_t^{(i)}=\frac{\alpha\|\bar{g}_t(X_i)\|_2^2}{2\sigma^2}$. Therefore, by properties of the individual filter, as long as $\frac{\alpha\sum_{j=1}^t\|\bar{g}_j(X_i)\|_2^2}{2\sigma^2}\leqslant B$, the output is (α,B)-individually RDP with respect to X_i. The clipping step ensures this inequality holds with $B=\frac{\alpha B_{\text{norm}}}{2\sigma^2}$ for all $t\in\mathbb{N}$ and for all data points X_i, and therefore the algorithm is $\left(\alpha,\frac{\alpha B_{\text{norm}}}{2\sigma^2}\right)$-RDP. ∎

When $B_{\text{norm}}=kC^2$, the privacy guarantees of Algorithm 8 are the same as those of Algorithm 7. However, they *do not* depend on the total number of steps $k_{\max}$ — in particular, $k_{\max}$ need not be equal to k. A natural question here is how to set the number of rounds $k_{\max}$ in Algorithm 8. (Certainly $k_{\max}$ should be at least $\lfloor B_{\text{norm}}/C^2\rfloor$, otherwise the privacy budget is not used up for any data point.) If $k_{\max}$ is relatively small, we might stop the optimization process too early, and thus forgo the possibility of achieving a higher accuracy. If, on the other hand, we set $k_{\max}$ to be too large, then a lot of points might get filtered out, in which case we add high amounts of noise relative to the number of active points.

One solution is to periodically estimate the number of active points in a privacy-preserving fashion. In particular, after round $\lfloor B_{\text{norm}}/C^2\rfloor$, the analyst can estimate the size of the active set $\{i:\sum_{j=1}^t\|\bar{g}_j(X_i)\|_2^2\leqslant B_{\text{norm}}\}$ (which is just a linear query) and use it to stop the computation. To reduce the privacy cost of such estimates one can use the continual monitoring technique [13] since each point is filtered out only once. Alternatively, if one only wants to ensure that the size of the active set exceeds some fixed threshold, one can use the sparse vector technique [12, 8] and thus incur an even smaller privacy loss due to adaptive stopping.

We also remark that, in practice, it is more common to use private SGD, rather than batch gradient descent. However, with individual privacy accounting, random subsampling of points would require computing gradients for *all* points at every step anyway (given that there is positive probability of every point participating at any step). As a result, running SGD would be no less computationally expensive. Nevertheless, gradient descent requires fewer steps and we observe that it achieves a better privacy-utility tradeoff, so it should be the preferred choice whenever computationally feasible.

We compare the performance of private gradient descent (Algorithm 7) and its generalization with filtering (Algorithm 8) in two settings; one, in which we train a non-convex deep learning model on MNIST [23], and another where we train a logistic regression classifier on the Adult dataset [6]. All reported average accuracies and deviations are estimated over 10 trials. In both cases, we fix target differential privacy parameters (ε, δ), and evaluate the test accuracy. We set $\delta = 10^{-5}$, and vary the value of ε. For every ε, all algorithm hyperparameters are first tuned to achieve high test accuracy with private gradient descent. For private gradient descent with filtering, to make the comparison as clear as possible, we adopt the same hyperparameters. In addition, we tune $k_{\max}$ on a separate and independent set of runs of our experiments.

6.1 Experiments on MNIST

We train a convolutional neural network with the same architecture as in the MNIST example of the Pytorch DP library: `https://github.com/facebookresearch/pytorch-dp`. Since we run batch gradient descent and not SGD, we tune all hyperparameters from scratch. We observe that private gradient descent achieves a better privacy-accuracy tradeoff than existing SGD baselines.

For small ε, we observe that private gradient descent with filtering, with carefully chosen number of steps $k_{\max} > B_{\text{norm}}/C^2$, achieves a noticeably higher accuracy than private GD with the same learning rate, noise scale, clip value, and $k = B_{\text{norm}}/C^2$. Below we summarize the achieved test accuracies.

MNIST accuracies		
ε	Private gradient descent	Private gradient descent with filtering
0.3	$(92.80 \pm 0.52)\%$	$(93.18 \pm 0.32)\%$
0.5	$(94.62 \pm 0.43)\%$	$(94.90 \pm 0.26)\%$
1.2	$(96.56 \pm 0.15)\%$	$(96.56 \pm 0.15)\%$

For $\varepsilon = 0.3$, we set $\sigma = 170$, $C = 10$, $\eta_t \equiv \eta = 0.2$, and $k = 104$ for private GD without filtering. To achieve the same privacy guarantees using private GD with individual filtering, we set $B_{\text{norm}} = kC^2 = 10400$, and we increase the total number of steps by 20%, i.e. $k_{\max} = 125 \approx 1.2k$. Given these parameters, private GD achieves test accuracy of $(92.80 \pm 0.52)\%$, while private GD with filtering achieves $(93.18 \pm 0.32)\%$.

For $\varepsilon = 0.5$, we set $\sigma = 130$, $C = 15$, $\eta_t \equiv \eta = 0.15$, and $k = 180$ for private GD without filtering. For GD with individual filtering, we set $B_{\text{norm}} = kC^2 = 40500$, and we increase the total number of steps by 10%, i.e. $k_{\max} = 1.1k = 198$. Given these parameters, private GD achieves test accuracy of $(94.62 \pm 0.43)\%$, while private GD with filtering achieves $(94.90 \pm 0.26)\%$.

For $\varepsilon = 1.2$, we set $\sigma = 88$, $C = 10$, $\eta_t \equiv \eta = 0.25$, and $k = 460$ for private GD without filtering. This parameter configuration achieves accuracy of $(96.56 \pm 0.15)\%$. When private GD achieves such high accuracies, we observe little benefit to individual filtering. This is due to the fact that the proportion of points filtered out right after round $\lfloor B_{\text{norm}}/C^2 \rfloor$ is comparable to the proportion of points yet misclassified, suggesting that few misclassified points remain in the active pool. Therefore, we set $B_{\text{norm}} = kC^2 = 46000$, and $k_{\max} = k$. Note that the accuracy achieved by private GD significantly exceeds the accuracy of $(94.63 \pm 0.34)\%$, previously reported for roughly the same ε (1.19), using the same architecture and private SGD (`https://github.com/facebookresearch/pytorch-dp/blob/master/examples/mnist_README.md`).

For $\varepsilon = 0.3$, we implement a Rényi odometer and observe the histogram of privacy losses at different training steps. Figure 2 shows the histogram of individual privacy odometers at step $k/2 = 52$, $k = 104$, and $k_{\max} = 125$. We set $\alpha = 63$ (which approximately optimizes the conversion from RDP to DP via Fact 2.3), and $\Delta = 0.00109 \approx \alpha/\sigma^2$. Our experiment suggests that the individual privacy odometers give

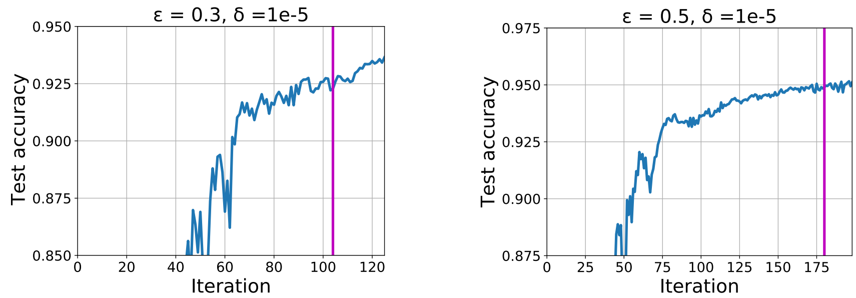

Figure 1: One run of private GD with filtering, for $\varepsilon = 0.3$ (left) and $\varepsilon = 0.5$ (right). For $\varepsilon = 0.3$, at step 104 (which is when private GD terminates) the test accuracy is 92.28%, while at step 125 (which is when private GD with filtering terminates) the test accuracy is 93.65%. For $\varepsilon = 0.5$, at step 180 (which is when private GD terminates) the test accuracy is 94.90%, while at step 198 (which is when private GD with filtering terminates) the test accuracy is 95.11%.

a reasonably tight estimate of the privacy loss despite the discretization error — for example, in the rightmost plot, we observe that the highest individual odometer values are around 0.135, and RDP loss of 0.135 translates to $\varepsilon \approx 0.32$, thus slightly overestimating $\varepsilon = 0.3$.

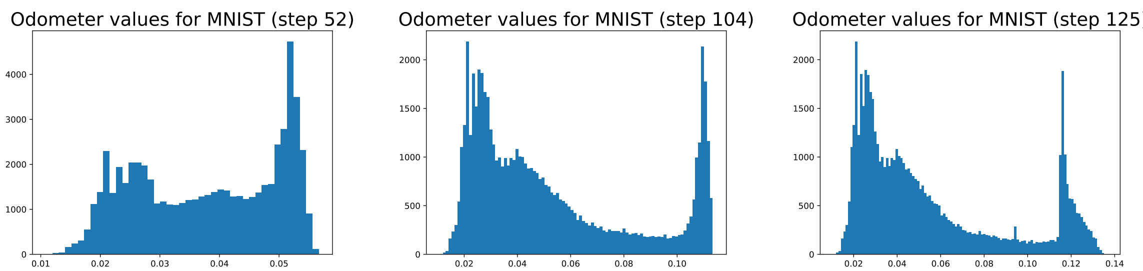

Figure 2: Histogram of individual privacy odometers at step $k/2 = 52$ (left), $k = 104$ (middle), and $k_{\max} = 125$ (right). The parameters are set as for $\varepsilon = 0.3$, and $\alpha = 63, \Delta = 0.00109$.

6.2 Experiments on Adult

We perform additional evaluations in a convex setting, by training a logistic regression classifier on the Adult dataset. This setting was also studied by Iyengar et al. [21], who report a non-private accuracy baseline of 84.8%. The table below summarizes our results.

Adult accuracies		
ε	Private gradient descent	Private gradient descent with filtering
0.3	$(83.80 \pm 0.22)\%$	$(83.91 \pm 0.12)\%$
0.5	$(84.11 \pm 0.15)\%$	$(84.18 \pm 0.10)\%$
1.0	$(84.28 \pm 0.11)\%$	$(84.42 \pm 0.12)\%$
1.2	$(84.45 \pm 0.13)\%$	$(84.48 \pm 0.15)\%$

For $\varepsilon = 0.3$, we set $\sigma = 455.34$, $C = 3.70$, $\eta_t \equiv \eta = 1.5$, and $k = 800$ for private GD without filtering. For private GD with individual filtering, we set $B_{\text{norm}} = kC^2$, and we increase the total number of steps by 20%, i.e. $k_{\max} = 960 = 1.2k$. Given these parameters, private GD achieves test accuracy of $(83.80 \pm 0.22)\%$, while private GD with filtering achieves $(83.91 \pm 0.12)\%$.

For $\varepsilon = 0.5$, we set $\sigma = 433.80$, $C = 3.70$, $\eta_t \equiv \eta = 1.5$, and $k = 2000$ for private GD without filtering. For GD with individual filtering, we set $B_{\text{norm}} = kC^2$, and we increase the total number of steps by 5%, i.e. $k_{\max} = 1.05k = 2100$. Given these parameters, private GD achieves test accuracy of $(84.11 \pm 0.15)\%$, while private GD with filtering achieves $(84.18 \pm 0.10)\%$.

For $\varepsilon = 1.0$, we set $\sigma = 613.49$, $C = 3.70$, $\eta_t \equiv \eta = 2$, and $k = 4000$ for private GD without filtering. This parameter configuration achieves accuracy of $(84.28 \pm 0.11)\%$. For private GD with individual filtering, we set $B_{\text{norm}} = kC^2$, and we increase the total number of steps by 20%, i.e. $k_{\max} = 1.2k = 4800$. This results in accuracy $(84.42 \pm 0.12)\%$.

For $\varepsilon = 1.2$, we set $\sigma = 259.33$, $C = 3.70$, $\eta_t \equiv \eta = 2$, and $k = 4000$ for private GD without filtering, thus achieving accuracy of $(84.45 \pm 0.13)\%$. For private GD with individual filtering, we set $B_{\text{norm}} = kC^2$, and we increase the total number of steps by 3%, i.e. $k_{\max} = 1.03k = 4120$. This results in accuracy $(84.48 \pm 0.15)\%$.

Acknowledgements

We thank Katrina Ligett, Kunal Talwar, and Neil Vexler for insightful discussions on individual notions of privacy and feedback on this work. We are grateful to Ryan McKenna for providing code and suggestions for the convex optimization experiments.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[2] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

[3] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[4] Rachel Cummings and David Durfee. Individual sensitivity preprocessing for data privacy. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 528–547. SIAM, 2020.

[5] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[6] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[7] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[8] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[9] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[10] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT*, pages 486–503, 2006.

[11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pages 265–284, 2006. URL `http://dx.doi.org/10.1007/11681878_14`.

[12] Cynthia Dwork, Moni Naor, Omer Reingold, Guy N Rothblum, and Salil Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390, 2009.

[13] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N Rothblum. Differential privacy under continual observation. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 715–724, 2010.

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

[15] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015.

[16] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

[17] Hamid Ebadi, David Sands, and Gerardo Schneider. Differential privacy: Now it's getting personal. *Acm Sigplan Notices*, 50(1):69–81, 2015.

[18] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[19] Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pages 535–544, 2018.

[20] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 199–208, 2011.

[21] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316. IEEE, 2019.

[22] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[23] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[24] Katrina Ligett, Charlotte Peale, and Omer Reingold. Bounded-leakage differential privacy. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

[25] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[26] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2016.

[27] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[28] Ryan M Rogers, Aaron Roth, Jonathan Ullman, and Salil Vadhan. Privacy odometers and filters: Pay-as-you-go composition. In *Advances in Neural Information Processing Systems*, pages 1921–1929, 2016.

[29] David M Sommer, Sebastian Meiser, and Esfandiar Mohammadi. Privacy loss classes: The central limit theorem in differential privacy. *Proceedings on privacy enhancing technologies*, 2019(2):245–269, 2019.

[30] Yu-Xiang Wang. Per-instance differential privacy. *Journal of Privacy and Confidentiality*, 9(1), 2019.

[31] Steven Wu, Aaron Roth, Katrina Ligett, Bo Waggoner, and Seth Neel. Accuracy first: Selecting a differential privacy level for accuracy-constrained erm. *Journal of Privacy and Confidentiality*, 9(2), 2019.