Cactus Mechanisms: Optimal Differential Privacy Mechanisms in the Large-Composition Regime

Wael Alghamdi*, Shahab Asoodeh‡, Flavio P. Calmon*, Oliver Kosut†, Lalitha Sankar†, and Fei Wei†

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834438

***Abstract*—Most differential privacy mechanisms are applied (i.e., composed) numerous times on sensitive data. We study the design of optimal differential privacy mechanisms in the limit of a large number of compositions. As a consequence of the law of large numbers, in this regime the best privacy mechanism is the one that minimizes the Kullback-Leibler divergence between the conditional output distributions of the mechanism given two different inputs. We formulate an optimization problem to minimize this divergence subject to a cost constraint on the noise. We first prove that additive mechanisms are optimal. Since the optimization problem is infinite dimensional, it cannot be solved directly; nevertheless, we quantize the problem to derive near-optimal additive mechanisms that we call "cactus mechanisms" due to their shape. We show that our quantization approach can be arbitrarily close to an optimal mechanism. Surprisingly, for quadratic cost, the Gaussian mechanism is strictly sub-optimal compared to this cactus mechanism. Finally, we provide numerical results which indicate that cactus mechanisms outperform Gaussian and Laplace mechanisms for a finite number of compositions.**

The full proofs can be found in the extended version at [1]. This paper is Part I in a pair of papers, where Part II is [2].

I. Introduction

Likelihood ratios are at the heart of most privacy metrics. Consider the problem of quantifying the privacy loss suffered by a sensitive variable X given an observation of a disclosed variable Y. For example, X may represent a dataset and Y a randomized function computed over X. Privacy can be measured in terms of properties of the *privacy loss random variable*, defined as

$$L_{x,x'} \coloneqq \log \frac{dP_{Y|X=x}}{dP_{Y|X=x'}}(Y), \tag{1}$$

where $Y \sim P_{Y|X=x}$ and $x, x' \in \mathcal{X} \coloneqq \mathsf{supp}(X)$. The channel $P_{Y|X}$ is often referred to as a *privacy mechanism*.

Today, the most popular privacy definition (including, in practice [3]–[5]) is *differential privacy* (DP), which quantifies privacy in terms of $L_{x,x'}$ when x, x' are close or "neighboring." Thus, given a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $P_{Y|X}$ is said to be (ε, δ)-differentially private ((ε, δ)-DP) [6] if

$$\sup_{d(x,x')\le s} \sup_{A\subset\mathcal{Y}} \left[P_{Y|X=x}(A) - e^{\varepsilon} P_{Y|X=x'}(A)\right] \le \delta, \tag{2}$$

where s determines when inputs x and x' are neighboring, and $\mathcal{Y} := \mathsf{supp}(Y)$. Intuitively, if a mechanism is (ε, δ)-differentially private for sufficiently small ε and δ, then an adversary observing Y cannot accurately distinguish between small changes in X.

Most privacy mechanisms are applied several times on sensitive data. Quantifying privacy guarantees under multiple *compositions* of a mechanism is a challenging problem. In the simple case where the same mechanism $P_{Y|X}$ is independently applied n times on data X generating output Y^n, i.e., $P_{Y^n|X} = \prod_{i=1}^n P_{Y_i|X}$, the privacy loss random variable is given by

$$L^n_{x,x'} \coloneqq \sum_{i=1}^n \log \frac{dP_{Y_i|X=x}}{dP_{Y_i|X=x'}}(Y_i), \tag{3}$$

where $Y_i \sim P_{Y_i|X=x}$. Differential privacy can be cast in terms of the privacy loss random variable. The reader can directly verify that n independent applications of a mechanism $P_{Y|X}$ is (ε, δ)-DP if

$$\sup_{d(x,x')\le s} \mathbb{E}\left[\left(1 - e^{-(L^n_{x,x'}-\varepsilon)}\right)^+\right] \le \delta. \tag{4}$$

From the law of large numbers, the distribution of $L^n_{x,x'}/n$ will concentrate around its mean, the KL-divergence, as

$$\frac{1}{n}\mathbb{E}\left[L^n_{x,x'}\right] = D\left(P_{Y|X=x} \| P_{Y|X=x'}\right). \tag{5}$$

Since the function $f(u) \coloneqq (1-e^{-nu+\varepsilon})^+$ is non-decreasing, in the limit of large compositions, privacy mechanisms with lower values of $D(P_{Y|X=x}\|P_{Y|X=x'})$ will enjoy stronger (ε, δ)-DP guarantees. Thus, regardless of the exact distribution of the privacy loss random variable, its mean (5) plays a central role in the privacy guarantees offered after many compositions. In applications such as privacy-ensuring machine learning, the number of compositions frequently exceeds $n = 10^3$.

We study the design of privacy mechanisms with favorable (ε, δ)-DP guarantees under a large number of compositions. Our approach departs from previous work in that we focus on the large-composition regime instead of optimizing (2). Since after many compositions, privacy will be mostly determined by the mean of the privacy loss random variable (5), we solve

*Corresponding author, remaining authors in alphabetical order. *W. Alghamdi and F. P. Calmon are with the School of Engineering and Applied Science, Harvard University (emails: alghamdi@g.harvard.edu, flavio@seas.harvard.edu)

‡ S. Asoodeh is with the Department of Computing and Software, McMaster University (email: asoodehs@mcmaster.ca)

† O. Kosut, L. Sankar, and F. Wei are with the School of Electrical, Computer, and Energy Engineering, Arizona State University (emails: {okosut,lsankar,fwei16}@asu.edu)

This material is based upon work supported by the National Science Foundation under Grant Nos. CAREER-1845852, CIF-1900750, CIF-1815361, CIF-1901243, CIF-1908725, CIF-2007688, and SaTC 2031799. The authors also thank Oracle Research for a gift that supported this work.

the optimization problem given in

$$\begin{aligned} \inf_{P_{Y|X}\in\mathscr{R}} \quad & \sup_{|x-x'|\le s} D(P_{Y|X=x}\|P_{Y|X=x'}) \\ \text{subject to} \quad & \sup_{x\in\mathbb{R}} \mathbb{E}[c(Y-x)\mid X=x]\le C, \end{aligned} \tag{6}$$

where $c:\mathbb{R}\to[0,\infty)$ is a pre-specified cost function, $s, C>0$ are constants, and $\mathscr{R}$ is the set of all Markov kernels on $\mathbb{R}$. Note that the cost function is critical: without the constraint, (6) can be trivially solved by any mechanism that is independent of X.

Our main contributions are as follows:

1) We show (Thm. 1) that additive mechanisms—i.e., where $Y = X + Z$ for a noise variable Z independent of X—suffice to minimize (6).
2) Even restricting to additive mechanisms, (6) is an infinite-dimensional optimization problem, so it cannot be solved directly. Instead, we formulate an approximate problem that is finite dimensional and can be solved efficiently. We prove (Thm. 3) that this approximate problem can get arbitrary close to optimal.
3) We solve the approximate problem to derive (near) optimal mechanisms for the quadratic cost function, i.e., $c(x)=x^2$. We dub the resulting mechanism the "cactus mechanism" due to the shape of the distribution (see Fig. 1). Surprisingly, the Gaussian distribution is strictly sub-optimal for (6), as the cactus mechanism achieves a smaller KL divergence for the same variance.
4) We bound the (ε,δ)-DP for the cactus mechanism in the context of sub-sampled stochastic gradient descent using the moments accountant method. Compared to the same analysis applied to a Gaussian mechanism, our approach does better for a reasonable number of compositions.

A. Related Work

Identifying optimal mechanisms is a fundamental and challenging problem in the domain of differential privacy. There have been several works in the literature that have attempted to address this problem. For instance, within the class of additive noise mechanisms and under the single shot setting (i.e., no composition), Ghosh et al. [7] showed that the geometric mechanism is universally optimal for $(\varepsilon,0)$-DP in a Bayesian framework, and Gupte and Sundararajan [8] derived the optimal noise distribution in a minimax cost framework. For a rather general cost function, the optimal noise distribution was shown to have a staircase-shaped density function [9]–[11]

Geng and Viswanath [12] showed that for (ε,δ)-DP and integer-valued query functions, in the single-shot setting, the discrete uniform noise distribution and the discrete Laplacian noise distribution are asymptotically optimal (for L^1 and L^2 costs) within a constant multiplicative gap in the high privacy regime (i.e., both ε and δ approach zero). Geng et al. [13] studied the same setting except for real-valued query functions and identified truncated Laplace distribution is asymptotically optimal in various high privacy regimes. Finally, Geng et al. [14] showed that the optimal noise distribution for real-valued query and $(0,\delta)$-DP is uniform with probability mass at the origin. Our work differs from these works in that we focus on the optimal mechanisms under a large number of compositions, rather than the single shot setting.

When considering a composition of n mechanisms, an important line of research has been to derive tighter composition results: relationships between the DP parameters of the composed mechanism and the parameters of each constituent mechanism. There are several composition results in the literature, such as [15]–[20]. More recently, Dong et al. [21] have proposed a composition result for large n and for a new variant of DP, called Gaussian-DP, that leverages the central limit theorem. These results can be sub-optimal (see, for example, [22, Fig. 1]). Consequently, numerical composition results have gained increasing traction as they lead to easier, yet powerful, methods for accounting the privacy loss in composition [22]–[25]. In particular, Koskela et al. [23] obtained a numerical composition result based on a numerical approximation of an integral that gives the DP parameters of the composed mechanism. The approximation is carried out by discretizing the integral and by evaluating discrete convolutions via the fast Fourier transform algorithm. The running time and memory needed for this approximation were subsequently improved [22]. While our work shares the focus on the large composition regime, we are primarily interested in synthesizing optimal mechanisms rather than analyzing existing mechanisms.

B. Notation

The Lebesgue measure on $\mathbb{R}$ is denoted by λ. We denote by $\mathscr{R}$ the set of all Markov kernels[1] on $\mathbb{R}$, i.e., conditional distributions $P_{Y|X}$ for $\mathbb{R}$-valued X and Y such that $x\mapsto P_{Y|X=x}(B)$ is a Borel function for all Borel sets $B\subset\mathbb{R}$. The set $\mathscr{B}$ denotes all Borel probability measures on $\mathbb{R}$. We fix a real-valued random variable X throughout, and let $P_X\in\mathscr{B}$ be its induced Borel probability measure. The KL-divergence is denoted by $D(P\|Q)$, and also by $D(p\|q)$ if $P,Q\ll\lambda$ with densities p and q. The expectation is denoted by $\mathbb{E}_P[f]:=\int_{\mathbb{R}} f\,dP$, and also by $\mathbb{E}_p[f]$ if $P\ll\lambda$ has probability density function (PDF) p. We let T_a denote the shift operator, i.e., for a function f of a real variable the function $T_a f$ is defined as $(T_a f)(x):=f(x-a)$, and for a measure P the measure $T_a P$ is defined by $(T_a P)(B):=P(B-a)$.

II. Optimality of Additive Continuous Channels

We start by deriving characterizations of solutions to the optimization problem (6). The difficulty of this problem lies in the fact that we are optimizing over all conditional distributions. This not only makes the problem infinite-dimensional, but it also renders direct approaches ineffective. The main result of this section, shown in Theorem 1, is that it suffices to consider continuous additive channels. In other words, the optimization in (6) may be restricted to conditional distributions of the form $P_{Y|X=x}=T_xP$ for some Borel probability measure P on $\mathbb{R}$ that is absolutely continuous with respect to the Lebesgue measure. Equipped with this reduction, we build in the next section an explicit family of finitely-parametrized distributions that are also optimal in (6).

[1]It is true that any conditional distribution from $\mathbb{R}$ into $\mathbb{R}$ has a version that is a Markov kernel [26, Chapter 4, Theorem 2.10].

A. Assumptions and Definitions

Throughout the paper, we require the cost function to satisfy the following properties.

Assumption 1. *The cost function $c : \mathbb{R} \to \mathbb{R}$ satisfies:*

- Positivity: $c(x) \geq 0$ *for all* $x \in \mathbb{R}$, *and* $c(0) = 0$.
- Symmetry: $c(x) = c(-x)$ *for all* $x \in \mathbb{R}$.
- Monotonicity: $c(x) \leq c(x')$ *if* $|x| \leq |x'|$.
- Continuity: c *is continuous over* $\mathbb{R}$.
- Tail regularity: *There exist* $\alpha, \beta > 0$ *such that* $c(x) \sim \beta x^{\alpha}$ *as* $x \to \infty$.

A natural choice of cost function is the quadratic cost $c(x) = x^2$, but we allow $c(x)$ to be any function that satisfies the above assumptions. For example, $c(x) = |x|^{\alpha}$ for any positive α is a natural family of cost functions.

Let $\mathscr{P} \subset \mathscr{R}$ be the set of conditional distributions $P_{Y|X}$ satisfying the cost constraint in (6), i.e., set

$$\mathscr{P} := \left\{ P_{Y|X} \in \mathscr{R} \; ; \; \sup_{x \in \mathbb{R}} \mathbb{E}[c(Y - x) \mid X = x] \leq C \right\}. \tag{7}$$

The infimal value in (6) is then

$$\mathrm{KL}^{\star} := \inf_{P_{Y|X} \in \mathscr{P}} \sup_{x, x' \in \mathbb{R} : |x - x'| \leq s} D(P_{Y|X=x} \| P_{Y|X=x'}). \tag{8}$$

We are interested in computing $\mathrm{KL}^{\star}$, as well as mechanisms $P_{Y|X}$ that approach this optimal value. Note that, for clarity of presentation, we suppress the dependence on (s, c, C) in the notations $\mathscr{P}$ and $\mathrm{KL}^{\star}$.

In the main problem (6), we allow $P_{Y|X}$ to be any mechanism that produces Y given X. A more restrictive but natural and easy-to-implement class of mechanisms is the *additive* mechanism class. An additive mechanism is given by $P_{Y|X=x}(B) = T_x P(B)$ where P is a Borel probability measure on $\mathbb{R}$. In other words, an additive mechanism $P_{Y|X}$ has Y of the form $Y = X + Z$ for some noise random variable $Z \sim P \in \mathscr{B}$ that is independent of the input X. Let $\mathscr{P}_{\text{add}} \subset \mathscr{B}$ be the set of additive mechanisms satisfying the cost constraint in (6),

$$\mathscr{P}_{\text{add}} := \{ P \in \mathscr{B} \; ; \; \mathbb{E}_P[c] \leq C \}. \tag{9}$$

Since the KL-divergence is shift-invariant, restricting the optimization (6) to additive mechanisms amounts to considering the simplified optimization problem

$$\mathrm{KL}^{\star}_{\text{add}} := \inf_{P \in \mathscr{P}_{\text{add}}} \sup_{a \in \mathbb{R} : |a| \leq s} D(P \| T_a P). \tag{10}$$

Of course, it is immediate that $\mathrm{KL}^{\star} \leq \mathrm{KL}^{\star}_{\text{add}}$. In fact, we will show below that these quantities are the same, meaning that there is no loss in restricting to additive mechanisms.

B. Optimality of Continuous Additive Mechanisms

The optimization problem in (6) is a convex problem, but the fact that the feasible set $\mathscr{P}$ is of infinite dimension means it cannot be solved directly, nor do the tractable properties one expects of a convex optimization problem necessarily follow. For example, in any finite dimensional convex optimization problem, a symmetry in the problem leads to the same symmetry in the solution. In this problem, one can see that shifting the mechanism—i.e., given $P_{Y|X}$, construct $Q_{Y|X=x}(B) = P_{Y|X=x+z}(B + z)$ for some z—does not change the cost constraint nor the objective value in (6). Thus, one might be inclined to conclude that the optimal mechanism is invariant to a shift (i.e., is an additive mechanism). Unfortunately, the infinite-dimensional nature of the problem means that this conclusion is not immediate. We resolve this issue in the following theorem which states that additive mechanisms are in fact optimal in (6).

Theorem 1. *We have that*

$$\mathrm{KL}^{\star} = \mathrm{KL}^{\star}_{\text{add}}, \tag{11}$$

and there exists a $P^{\star} \in \mathscr{P}_{\text{add}}$ *achieving this value. Further, any such* $P^{\star}$ *is necessarily absolutely continuous.*

Proof sketch: The proof is given in the extended paper [1, Appendix A]. We give here only a high level description of the approach. Let $P^{(k)}_{Y|X}$ be a sequence achieving $\mathrm{KL}^{\star}$. We make these mechanisms increasingly closer to being additive, while sacrificing neither feasibility nor utility, by considering the convex combinations

$$\overline{P}^{(k)}_{Y|X=x}(B) := \mathbb{E}\left[P^{(k)}_{Y|X=x+Z_k}(B + Z_k) \right] \tag{12}$$

where $Z_k \sim \mathrm{Unif}([-k, k])$. Specifically, one can invoke Prokhorov's theorem on the $\overline{P}^{(k)}_{Y|X}$, thereby extracting a probability measure $P^{\star}$ such that $\overline{P}^{(k)}_{Y|X=x} \to T_x P^{\star}$ weakly for each fixed x. Finally, we show that the mechanism $P^{\star}$ is optimal by invoking joint convexity and lower-semicontinuity of the KL-divergence. ∎

Remark 1. The proof of $P^{\star} \ll \lambda$ only relies on the property that $P^{\star} \ll T_a P^{\star}$ for every $|a| \leq s$, which holds in view of $\mathrm{KL}^{\star} < \infty$. Therefore, any *feasible* additive mechanism must be absolutely continuous with respect to the Lebesgue measure, i.e., if $\mu \in \mathscr{B}$ satisfies $\sup_{|a| \leq s} D(\mu \| T_a \mu) < \infty$ then we necessarily have $\mu \ll \lambda$.

III. Numerical Approximation: The Cactus Distribution

The optimization problem over additive mechanisms in (10) is infinite-dimensional, so it cannot be solved numerically as-is, and it appears to have no closed-form solution for non-trivial cost functions. The lack of closed-form solution is true even for the simple case of $c(x) = x^2$: to our surprise, as will be illustrated later, the Gaussian mechanism is not optimal![2] In our companion paper [2], we explore the regime where $s \to 0^+$; in this limit, we show that the optimal distribution can be determined exactly, and in fact for quadratic cost the limiting optimal distribution *is* Gaussian—although for other costs the optimal distribution is much more surprising.

In the regime of fixed positive s, to find practically achievable near-optimal mechanisms, we resort to numerical approximation of (10). In this section, we fix $s = 1$. We can do

[2]Of course, simply because Gaussian is not optimal does not imply that there is no closed-form solution. It is possible to write a set of KKT conditions for (10), which we have omitted from this paper in the interest of space. This set of KKT conditions cannot be solved in closed-form.

this without loss of generality simply by scaling: that is, the optimization problem in (10) with sensitivity s and cost function $c(x)$ is equivalent to the same problem with sensitivity 1 and cost function $c(sx)$.

To approximate (10) by a numerically tractable problem, we (i) quantize the distribution, and (ii) only explicitly parameterize the distribution in a certain interval. Specifically, we construct a mapping from finite-length vectors to continuous distributions as follows.

Definition 1. Fix two positive integers n and N, and a constant $r \in (0,1)$. Consider the partition of $\mathbb{R}$ by intervals $\{\mathcal{J}_{n,i}\}_{i\in\mathbb{Z}}$ defined by: $\mathcal{J}_{n,0} := [-1/(2n), 1/(2n)]$ and

$$\mathcal{J}_{n,i} := \begin{cases} \left(\frac{i-1/2}{n}, \frac{i+1/2}{n}\right], & \text{if } i > 0, \\ \left[\frac{i-1/2}{n}, \frac{i+1/2}{n}\right), & \text{if } i < 0. \end{cases} \tag{13}$$

We associate to each vector $\boldsymbol{p} = (p_0, p_1, \ldots, p_N) \in [0,1]^{N+1}$ a piecewise constant function that is defined by

$$f_{n,r,\boldsymbol{p}}(x) = \begin{cases} np_{|i|}, & \text{if } x \in \mathcal{J}_{n,i}, \text{with } |i| < N, \\ np_N r^{|i|-N}, & \text{if } x \in \mathcal{J}_{n,i}, \text{with } |i| \geq N. \end{cases} \tag{14}$$

We also associate with $f_{n,r,\boldsymbol{p}}$ the Borel measure $P_{n,r,\boldsymbol{p}}$, where

$$P_{n,r,\boldsymbol{p}}(B) := \int_B f_{n,r,\boldsymbol{p}}(x)\, dx. \tag{15}$$

Remark 2. Note that

$$\int_{\mathbb{R}} f_{n,r,\boldsymbol{p}}(x)\, dx = p_0 + \sum_{i=1}^{N-1} 2p_i + \frac{2p_N}{1-r} =: S_{r,\boldsymbol{p}}. \tag{16}$$

If $S_{r,\boldsymbol{p}} = 1$, then $P_{n,r,\boldsymbol{p}}$ is a probability measure with density $f_{n,r,\boldsymbol{p}}$. This distribution is symmetric around the origin, i.e., $f_{n,r,\boldsymbol{p}}(x) = f_{n,r,\boldsymbol{p}}(-x)$. Further, its tails decay almost geometrically: for $(N+1/2)/n < x_1 < x_2$ one has $f_{n,r,\boldsymbol{p}}(x_2) = r^{nk} f_{n,r,\boldsymbol{p}}(x_1)$ where $k = (\lceil nx_2 - 1/2\rceil - \lceil nx_1 - 1/2\rceil)/n \approx x_2 - x_1$.

The main results of this section are: we show that the distribution family introduced in Definition 1 is optimal for (6), and we show that the optimal distribution *within* this family (which we will call the *cactus distribution*) is obtainable via a tractable finite-dimensional convex optimization problem.

We use the following notation. Consider the restriction of (10) to the mechanisms constructible by Definition 1. For a fixed triplet $(n, N, r) \in \mathbb{N}^2 \times (0,1)$, denote the set of such mechanisms by $\mathscr{C}_{n,N,r} \subset \mathscr{B}$, i.e.,

$$\mathscr{C}_{n,N,r} := \left\{ P_{n,r,\boldsymbol{p}} \;;\; \boldsymbol{p} \in [0,1]^{N+1}, S_{r,\boldsymbol{p}} = 1 \right\}. \tag{17}$$

(Recall the definition of $S_{r,\boldsymbol{p}}$ from (16).) Denote the optimal value achievable by the class $\mathscr{C}_{n,N,r}$ by

$$\mathrm{KL}^\star_{n,N,r}(C) := \inf_{\substack{P\in\mathscr{C}_{n,N,r} \\ \mathbb{E}_P[c]\leq C}} \sup_{|a|\leq 1} D(P\|T_a P). \tag{18}$$

We show next that we may restrict the shift a in (18) to take values over the finite set $\{1/n, 2/n, \cdots, 1\}$ (rather than varying over the whole interval $[-1,1]$), thereby rendering (18) a finite-dimensional optimization problem amenable to standard numerical convex-programming methods. For each $i \in \mathbb{Z}$, we denote the constants

$$c_{n,i} := \int_{\mathcal{J}_{n,i}} nc(x)\, dx. \tag{19}$$

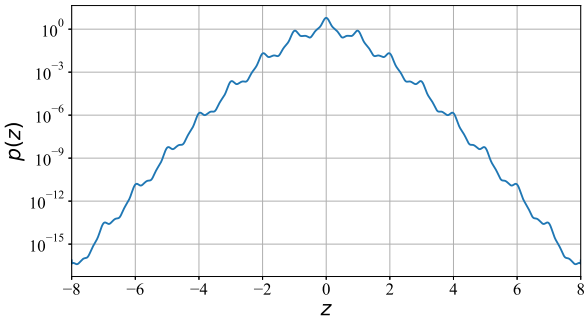


Fig. 1: The optimal distribution $p(z)$, found by solving (20) (and dubbed the *cactus distribution*), plotted on a semi-log scale. The cost function is $c(z) = z^2$, and the parameters are: $s = 1$, $C = 0.25$, $n = 200$, $N = 1600$, and $r = 0.9$.

Theorem 2. *Fix $r \in (0,1)$, and positive integers $n < N$. The minimization (18) can be recast as the following convex program over the variable $\boldsymbol{p} = (p_0, \cdots, p_N) \in \mathbb{R}^{N+1}$*

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}} \quad & \max_{k\in\{1,\ldots,n\}} \frac{1}{2} \sum_{i=-N+1}^{N-k-1} (p_{|i|} - p_{|i+k|}) \log \frac{p_{|i|}}{p_{|i+k|}} \\ & + \sum_{i=N-k}^{N-1} (p_i - p_N r^{i+k-N}) \log \frac{p_i}{p_N r^{i+k-N}} \\ & + p_N \frac{1-r^k}{1-r} k \log r^{-1} \\ \text{subject to} \quad & p_0 c_{n,0} + \sum_{i=1}^{N-1} 2p_i c_{n,i} + 2p_N \sum_{i=N}^{\infty} c_{n,i} r^{i-N} \leq C, \\ & p_0 + \sum_{i=1}^{N-1} 2p_i + \frac{2p_N}{1-r} = 1, \\ & p_i \geq 0 \text{ for all } i \in \{0, \ldots, N\}. \end{aligned} \tag{20}$$

Figure 1 shows an example of the distribution that results from the finite-dimensional optimization problem in (20) with a quadratic cost. The shape of this distribution[3] has inspired the name the "cactus distribution."

The following result shows that cactus mechanisms derived from the optimization problem (20) are in fact globally optimal for the main optimization problem (6).

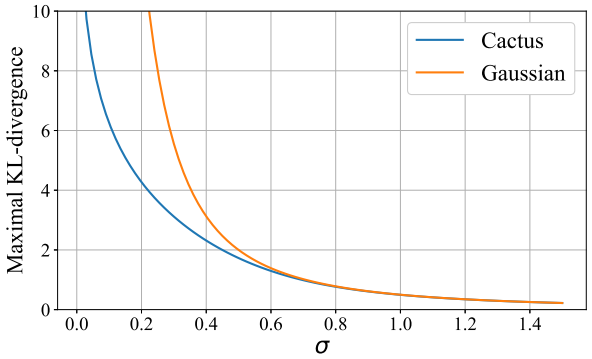
Theorem 3. *Denote the optimal value a cactus distribution can achieve by*

$$\mathrm{KL}^\star_{\text{Cactus}} := \lim_{\varepsilon\to 0^+} \inf_{(n,N,r)\in\mathbb{N}^2\times(0,1)} \mathrm{KL}^\star_{n,N,r}(C+\varepsilon). \tag{21}$$

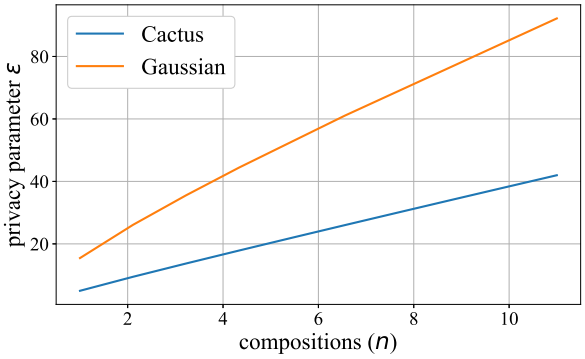
We have that $\mathrm{KL}^\star = \mathrm{KL}^\star_{\text{Cactus}}$.

Remark 3. The proof of Theorem 3 gives some guidelines for choosing the parameters (n, N, r). For example, optimal

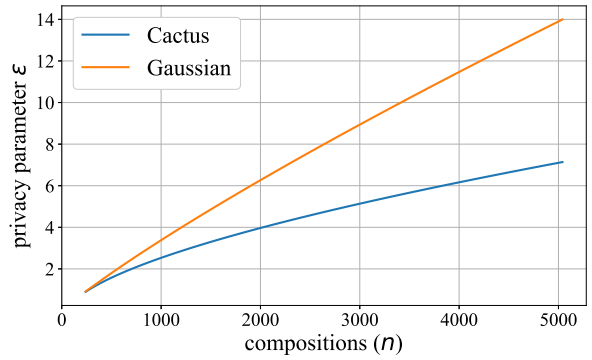
[3] In addition to the state of Arizona being home of several of the authors.



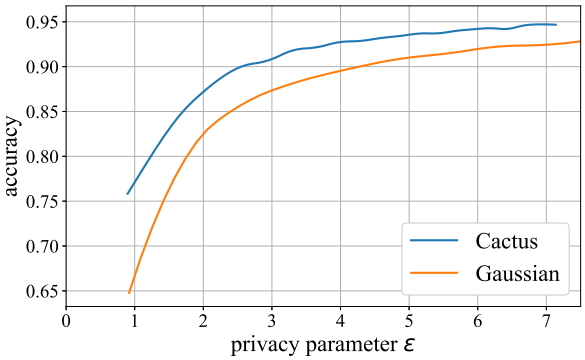
(a) Achieved maximal KL-divergence $\sup_{|a|\le s} D(p\|T_a p)$ versus σ, the (quadratic) cost constraint is of the form $\mathbb{E}[Z^2] \le \sigma^2 = C$ with fixed sensitivity $s = 1$.



(b) Privacy parameter ε versus the number of compositions, computed via the moments accountant, where $\delta = 10^{-3}$, and quadratic cost $C = 0.1$ with fixed sensitivity $s = 1$.



(c) Privacy parameter ε versus the number of compositions, computed via the moments accountant, where $\delta = 10^{-5}$, subsampling rate $q \approx 0.00417$, and quadratic cost $C = 0.1$ with fixed sensitivity $s = 1$.



(d) Model accuracy versus privacy parameter ε. The settings are the same as in Figure 2c and experiment details are given in Section IV.

Fig. 2: Comparison between the Gaussian and cactus mechanisms.

cactus distributions can be obtained by restricting the ratio N/n (chosen sufficiently large), and choosing $r = 1 - \Theta_\alpha(N^{-1})$.

IV. Numerical Results

We solve the optimization problem (20) using an interior-point method. An example of the cactus distribution for quadratic cost is shown in Figure 1. Figure 2a compares the maximal KL-divergence achieved by the cactus to that of Gaussian distributions for fixed sensitivity $s = 1$ and various σ. As noted above, varying σ with fixed s is equivalent to varying s with fixed σ. The KL-divergence for cactus is computed numerically, and for Gaussian mechanisms the KL-divergence is exactly $\frac{1}{2\sigma^2}$. The cactus distribution outperforms the Gaussian distribution in terms of KL-divergence for all values of σ, although the difference decreases as σ grows such that for larger values of σ it is difficult to discern any gap between the curves in Figure 2a. (Our companion paper [2] gives a theoretical explanation for why Gaussian is so close to optimal as s/σ decreases.) To illustrate that this improvement in KL-divergence leads to an improvement in (ε, δ)-DP, we compute the achieved privacy via moments accountant [18] for each mechanism. Figure 2b shows the resulting ε value as a function of the number of compositions, for fixed $\delta = 10^{-3}$. Indeed, the cactus mechanism does better than Gaussian.

To give a reasonable comparison in the context of machine learning, we modified the tutorial code in TensorFlow-Privacy [27], which implements the DP-stochastic gradient descent (SGD) algorithm with a Gaussian mechanism on a convolutional neural network (CNN) model. We use the training results from the original tutorial as a benchmark, then replace the Gaussian mechanism with our cactus mechanism, and train the model using the renewed setting. We select a noise level $\sigma = \sqrt{0.1}$. We test the original and modified model on a popular image dataset, MNIST, which is of size 60000. We choose a batch-size 250, such that each epoch consists of 240 iterations (i.e., compositions) and the sub-sampling rate[4] is $q = 250/60000 \approx 0.00417$. Figure 2c shows the achieved (ε, δ)-DP as computed by the moments account in this setting. Fixing $\delta = 10^{-5}$, Figure 2d shows the tradeoff between privacy ε and accuracy of the resulting CNN as the number of training iterations increases. One can see that for a fixed privacy budget (i.e., fixed ε and δ), the cactus mechanism allows more training iterations and, thus, better accuracy.

[4]The cactus mechanism is not optimized for subsampling. Nevertheless, we observe numerical performance of the cactus mechanism in the subsampling setting outperforming that of the Gaussian mechanism.

REFERENCES

[1] W. Alghamdi, S. Asoodeh, F. Calmon, O. Kosut, L. Sankar, and F. Wei, "Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime," 2022. [Online]. Available: https://github.com/WaelAlghamdi/DP-Cactus

[2] ——, "Schrödinger mechanisms: Optimal differential privacy mechanisms for small sensitivity," 2022. [Online]. Available: https://github.com/WaelAlghamdi/DP-Schrodinger

[3] Ú. Erlingsson, V. Pihur, and A. Korolova, "Rappor: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*. ACM, 2014, pp. 1054–1067.

[4] Differential privacy team Apple, "Learning with privacy at scale," 2017.

[5] D. Kifer, S. Messing, A. Roth, A. Thakurta, and D. Zhang, "Guidelines for implementing and auditing differentially private systems," *ArXiv*, vol. abs/2002.04049, 2020.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory of Cryptography (TCC)*, Berlin, Heidelberg, 2006, pp. 265–284.

[7] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," *SIAM Journal on Computing*, vol. 41, no. 6, pp. 1673–1693, 2012.

[8] M. Gupte and M. Sundararajan, "Universally optimal privacy mechanisms for minimax agents," in *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2010, p. 135–146.

[9] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 925–951, 2015.

[10] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1176–1184, 2015.

[11] J. Soria-Comas and J. Domingo-Ferrer, "Optimal data-independent noise for differential privacy," *Information Sciences*, vol. 250, no. Complete, pp. 200–214, 2013.

[12] Q. Geng and P. Viswanath, "Optimal noise adding mechanisms for approximate differential privacy," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 952–969, 2016.

[13] Q. Geng, W. Ding, R. Guo, and S. Kumar, "Tight analysis of privacy and utility tradeoff in approximate differential privacy," in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108, 2020, pp. 89–99.

[14] ——, "Optimal noise-adding mechanism in additive differential privacy," in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and M. Sugiyama, Eds., vol. 89. PMLR, 16–18 Apr 2019, pp. 11–20. [Online]. Available: https://proceedings.mlr.press/v89/geng19a.html

[15] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 51–60.

[16] J. Murtagh and S. Vadhan, "The complexity of computing the optimal composition of differential privacy," in *Proc. Int. Conf. Theory of Cryptography*, 2016, pp. 157–175.

[17] P. Kairouz, S. Oh, and P. Viswanath, "The composition theorem for differential privacy," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., vol. 37, 2015, pp. 1376–1385.

[18] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[19] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut, and L. Sankar, "Three variants of differential privacy: Lossless conversion and applications," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 1, pp. 208–222, 2021.

[20] S. Meiser and E. Mohammadi, "Tight on budget? tight bounds for r-fold approximate differential privacy," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18, 2018, p. 247–264.

[21] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *arXiv preprint arXiv:1905.02383*, 2019.

[22] S. Gopi, Y. T. Lee, and L. Wutschitz, "Numerical composition of differential privacy," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[23] A. Koskela, J. Jälkö, and A. Honkela, "Computing tight differential privacy guarantees using fft," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2560–2569.

[24] A. Koskela, J. Jälkö, L. Prediger, and A. Honkela, "Tight differential privacy for discrete-valued mechanisms and for the subsampled gaussian mechanism using fft," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Banerjee and K. Fukumizu, Eds., vol. 130. PMLR, 13–15 Apr 2021, pp. 3358–3366. [Online]. Available: https://proceedings.mlr.press/v130/koskela21a.html

[25] Y. Zhu, J. Dong, and Y.-X. Wang, "Optimal accounting of differential privacy via characteristic function," *arXiv preprint arXiv:2106.08567*, 2021.

[26] E. Çınlar, *Probability and Stochastics*. New York, NY: Springer, 2011.

[27] TensorFlow-Privacy tutorial, https://github.com/tensorflow/privacy.git/.