Private Learning of Linear Orders

L.C. Brown

April 2024

1 Introduction

A measure of privacy is necessary for individual liberty, and most members of industrial societies have noticed that the increase in information storage capacity over the past twenty years has eliminated not only the urban anonymity many were accustomed to but also the little privacy one may assume was enjoyed even in early agrarian societies. Some learning theorists have addressed the problem of leveraging the enormous power of machine learning without compromising human privacy; here, we examine in some detail a simple theoretical learning case which illuminates techniques and subtle problems we hope future work will explore in more complex cases.

Our mathematical framework is based on the definitions of Dwork and her collaborators in [4], where the theory of differential privacy is introduced as a useful mathematical model for the intuitive concept of privacy. Their arguments for their philosophy, besides a wealth of mathematical analysis, can be found in [7]. A practical implementation of a differentially private learning algorithm can be found in [1]. Many organizations, including Google and the United States Census Bureau, have used the theory of differential privacy or variants of it [8] [5].

Differential privacy is a desirable property of a learning algorithm for many reasons: the name promises protection from governments, companies, and other organizations with large information storage capacity, but the mathematical definition also guarantees some degree of immunity to outlier data and resistance to measurement error. Even an unscrupulous actor would prefer to use a differentially private algorithm if it were otherwise equivalent to a less private algorithm.

We consider the problem of learning upwards-closed subsets of the real unit interval $[0, 1]$. The non-existence of a differentially private learner which learns this set probably approximately correctly for every measure on its natural Borel $\sigma-$algebra was shown in [3], and the result was extended to impurely differentially private learners by [2]. However, there are cases where the measure on the sample space is known, at least approximately, and there the requirement that the learner be correct for every measure is unnecessary. When examining, say, the relationship between human height and liver failure, one need not find a learner which would learn well even if every human in the training data were exactly 175 cm or 195 cm tall. By demonstrating the significance of knowing the measure in our case, which is the simplest case where the results of [2] imply there is no private learner which learns well with respect to every measure, we hope to inspire a general examination of the amount of information one needs about the measure when learning a given hypothesis class in order to maintain privacy without losing accuracy.[1]

We reproduce the definitions of differentially private learners and exponential mechanisms relevant to our purpose here in the first section, as we require definitions more precise than those extant in the literature. In the second section, we introduce our learner for the uniform measure by defining it on a certain subset of possible training data and demonstrating it can be interpolated to all training data, regardless of whether those data are realized by any member of our hypothesis class. The third section contains calculations which show our learners are probably approximately correct with respect to the uniform measure. We draw attention to the use of logistic functions as exponential mechanisms, as they are often overlooked in favor of Laplacians, which have a slightly simpler form but, for a fixed accuracy requirement, sacrifice privacy in the most likely cases in favor of privacy in unlikely cases. Our fourth section considers the case where the measure is non-uniform.

[1] Some accuracy must be lost, but one may hope that there is at least asymptotic agreement as the sample size increases.

2 Learners

Fix a set M and $\mathfrak{X} \subset \mathcal{P}(M)$ a collection of subsets of M. We shall refer to $\mathfrak{X}$ as our hypothesis class.

We also fix a σ-algebra $\mathcal{M}$ on M such that $\mathfrak{X} \subset \mathcal{M}$. I will denote the unit interval $[0,1] \subset \mathbf{R}$. We write 2 for the set $\{0,1\}$; we shall not distinguish between subsets of a given set A and functions from A to 2, or between a particular subset $B \subset A$ and its characteristic function $\chi_B : A \to 2$, where $\chi_B(a) = 1$ precisely when $a \in B$. When there is no danger of confusion, we write a for the one-element set $\{a\}$. The natural numbers are $\omega = \{0,1,2,...\}$, and $A^{<\omega} = \cup_{n=0}^{\infty} A^n$.

Let $\mathcal{N}$ be a σ-algebra on $\mathfrak{X}$ (that is, a collection of subsets of $\mathfrak{X}$) such that the evaluation map $M \times \mathfrak{X} \to 2$ which sends (m,x) to $x(m)$ is measurable with respect to the product σ-algebra $\mathcal{M} \times \mathcal{N}$. Let $\mathfrak{D}$ be the set of probability measures on $\mathcal{N}$.

What we call learners are sometimes called random learners, or randomized learners. One may think of elements of $\mathfrak{D}$ as being instantiated in some way so that some element of $\mathfrak{X}$ is ultimately "chosen" according to the measure; the definition of differential privacy says changing one training datum cannot greatly change the likelihood of any individual $X \in \mathfrak{X}$ being chosen; to be more precise, the likelihood of choosing an X in some fixed $A \subset \mathfrak{X}$ cannot change greatly.

Definition 1. (i) A learner is a map $\Lambda : (M \times 2)^{<\omega} \to \mathfrak{D}$. Let $\overline{M}$ denote the set of functions from M to I. The composition of a learner with the map $p : \mathfrak{D} \to \overline{M}$ which sends $\mu \in \mathfrak{D}$ to the function which sends x to $\int_{\mathfrak{X}} y(x) d\mu(y)$ is that learner's prediction.

(ii) A learner Λ is (ℓ, α)-differentially private if ℓ is a positive integer, $\alpha, \beta \geq 0$ are real, and whenever $A \in \mathcal{N}$ and $\nu = (\nu', \nu'')$, $\rho = (\rho', \rho'')$ are elements of $(M \times 2)^{\ell}$ such that there is exactly one $j < \ell$ where $\nu(j) \neq \rho(j)$ (that is, $\nu'(j) \neq \rho'(j)$ or $\nu''(j) \neq \rho''(j)$, the 'or' here as elsewhere being inclusive), we must have $\Lambda(\nu)(A) \leq e^{\alpha} \Lambda(\rho)(A)$.

(iii) A function $\Lambda : (M \times 2)^{\ell} \to \mathfrak{D}$ is said to be α-private if it satisfies the above definitions with the domain restricted to $(M \times 2)^{\ell}$.

Endless variations on the above definitions are possible; what we call differential privacy is sometimes called pure differential privacy. The essential point in the definition of differential privacy is that no restriction is placed on ν or ρ; neither $\mathfrak{X}$ nor $\mathcal{M}$ enter into (ii). The ν and ρ could be unrealizable or unlikely training data. This is a formalization of the definition in [7], and it is a natural definition from the conservative perspective of differential privacy, which demands the protection of individual privacy even in cases where the data analyst makes poor hypotheses or the collected data are inaccurate. In our setting, even one inaccurate datum may prevent our hypothesis class $\mathfrak{X}$ from realizing the training data, and dropping the privacy requirements in such cases amounts to an admission of defeat in all scientific applications.

Combinatorial conditions on probably approximately correct (PAC) learnability date to [6]; the combinatorial conditions necessary for PAC differentially private learning were established in [2].[2] Of course, private learners are easy to find, and the problem only becomes nontrivial when conjoined with the accuracy condition. Note that the accuracy of a learner depends only on its prediction, whereas the definition of differential privacy makes direct reference to the learner.

While we assume familarity with probably approximately correct learning, we reproduce a standard definition here to fix notation.

Definition 2. A learner Λ is (ℓ, ϵ, δ)–probably approximately correct (PAC) for the measure μ, where ℓ is a natural number, $\epsilon, \delta > 0$, if

$$\mu^{\ell}(\{\xi \in M^{\ell} : \mathcal{A}_{\mu_x}(\Lambda(\xi, x \circ \xi)) > 1 - \epsilon\}) > 1 - \delta.$$

We say Λ is (ℓ, ϵ, δ)–PAC if it is (ℓ, ϵ, δ)–PAC for every measure μ on $\mathcal{M}$.

The following definition is based on Chapter 3 of [7] and, in particular, Definition 3.3 of the same, where the Laplace mechanism is considered.

[2]The reader unfamiliar with PAC learning may also refer to this paper for the definition of PAC learning we adopt.

Definition 3. An exponential mechanism with rate at most $C > 0$ is a differentiable function $f : \mathbf{R} \to \mathbf{R}$ with $f, f' \geq 0$ such that, for all $a, x \in \mathbf{R}$, $f'(x+a) \leq e^{C|a|} f'(x)$. If f is an exponential mechanism, the rate of f is the least C such that f has rate at most C.

Lemma 1. *All functions of the form $e^{C(x-x_0)}$ or $(1+e^{-C(x-x_0)})^{-1}$, where $x_0 \in \mathbf{R}$ and $C \geq 0$, are exponential mechanisms; so are all antiderivatives of functions of the form $e^{C|x-x_0|}$, with x_0, C as above; in all cases, the rate of the exponential mechanism is the respective C.*

When we refer to logistic or Laplacian exponential mechanisms, we mean exponential mechanisms of the second and third types, respectively, and we shall take $x_0 = 0$ unless otherwise noted; further, we scale the Laplacian exponential mechanisms by 2^{-1} so that their integrals over the real line equal 1.

The fundamental theorem of calculus yields

Lemma 2. *If f is an exponential mechanism with rate C, then, for any real numbers a, b, and c, we have that $f(c+a) - f(c) \leq e^{C|b|}(f(c+a+b) - f(c+b))$.*

3 Privacy by Interpolation

We now describe how exponential mechanisms give private learners where $\mathfrak{X} = \{(a, 1] | a \in I\}$. To simplify notation, we identify elements of $\mathfrak{X}$ with their infima.[3] Consider $(I \times 2)^{<\omega} = \cup_{i<\omega}(I \times 2)^n$ as a graph where there is an edge between η and ν, written $\eta E \nu$, if there is exactly one $i < |\eta| = |\nu|$ such that $\eta(i) \neq \nu(i)$. Note that the graph has infinitely many connected components, as each sequence is only connected to other sequences of the same length.

Lemma 3. *Let $u : (I \times 2)^{<\omega} \to I$ and $d : \omega \to \mathbf{R}$ be functions such that, whenever $\eta E \nu$ in the aforementioned graph structure, $|u(\eta) - u(\nu)| \leq d(|\eta|)$, and, for each natural number i, let f_i be a strictly increasing exponential mechanism with rate $k(i)$. Then the learner $\Lambda_{u,f}$ which assigns to $\eta \in (I \times 2)^i$ the measure on $\mathfrak{X}$ with probability density function $\Theta(t) = \frac{f_i'(t-u(\eta))}{f_i(1-u(\eta)) - f_i(-u(\eta))}$ is purely $(i, 2k(i)d(i))$-differentially private for every i.*

Proof. This follows from Definition 3 and Lemmas 1 and 2. †

Lemma 4. *Let $r : \omega \to \mathbf{R}^+$, and take $S_r \subset (I \times 2)^{<\omega}$ to consist of those finite sequences (σ, τ) such that $\cup_{i<|\sigma|} B(r(|\sigma|), \sigma(i)) = I$, where $B(s, x) \subset I$ denotes the ball of radius s centered at x, and $\sigma(i) \leq \sigma(j) \implies \tau(i) \leq \tau(j)$. For $(\sigma, \tau) \in S_r$, put $u(\sigma, \tau) = 2^{-1}(\max_{i \in \tau^{-1}(0)} \sigma(i) + \min_{i \in \tau^{-1}(1)} \sigma(i))$, where we consider $\max \varnothing = 0, \min \varnothing = 1$. If there is a path of length m in $(I \times 2)^{<\omega}$ between $\nu, \eta \in S_r$, $|u(\nu) - u(\eta)| \leq 2mr(|\nu|)$.*

Proof. Let $\ell = |\eta| = |\nu|$. Assume without loss of generality that ν is weakly increasing and $u(\nu) < u(\eta)$. Further write $\nu = (\sigma, \tau)$. We may assume $\tau^{-1}(1) \neq \varnothing$. Pick i_0 maximal so that $\tau(i_0) = 0$; if $\tau^{-1}(0) = \varnothing$, let $i_0 = -1$. By assumption, there is some j with $i_0 < j \leq i_0 + m + 1$ such that $\nu(j) = \eta(j)$; take j_0 to be the least such j. Because $\nu \in S_r$ and is weakly increasing, the distance between $\sigma(i)$ and $\sigma(i+1)$ is less than $2r(\ell)$ for any i; consequently, $\sigma(j_0) < \sigma(i_0 + 1) + 2mr(\ell)$, and with $\eta = (\sigma', \tau')$,

$$\min_{i \in \tau'^{-1}(1)} \sigma'(i) \leq \sigma'(j_0) = \sigma(j_0).$$

Further, if $j_0 \leq i_0 + m$, then $\sigma(j_0) < \sigma(i_0) + 2mr(\ell)$. Since $\tau'(j_0) = \tau(j_0) = 1$ and $\sigma'(i) \leq \sigma'(j) \implies \tau'(i) \leq \tau'(j)$, we have $\max_{i \in \tau'^{-1}(0)} \sigma'(i) \leq \min_{i \in \tau^{-1}(1)} \sigma'(i) \leq \sigma(j_0)$. In this case

$$\begin{aligned}
\max_{i \in \tau'^{-1}(0)} \sigma'(i) + \min_{i \in \tau^{-1}(1)} \sigma'(i) &\leq 2 \min_{i \in \tau^{-1}(1)} \sigma'(i) \\
&\leq 2\sigma(j_0) \\
&< 2(\sigma(i_0) + 2mr(\ell)) \\
&\leq \sigma(i_0) + \sigma(i_0 + 1) + 4mr(\ell),
\end{aligned}$$

whence the result follows immediately. If $j_0 = i_0 + m + 1$, the only paths from ν to η of length m pass only through elements of S_r which agree with ν on $i \leq i_0$ and $i \geq j_0$. Consequently there are at most m i's such that

[3] The choice between open and closed intervals is arbitrary and, from the perspective of accuracy for uniform Lebesgue measure, one could even mix the two; the situation with non-uniform measures is more complicated.

$\sigma'(i_0) < \sigma'(i) < \sigma'(j_0)$, and we can conclude from the definition of S_r that, for any such i, $\sigma'(i) \le \sigma(i_0) + 2mr$. Thus $\max_{i \in \tau'^{-1}(0)} \sigma'(i) \le \sigma(i_0) + 2mr$; combining this with the bound $\sigma'(j_0) < \sigma(i_0 + 1) + 2mr(\ell)$ yields the result. †

Lemma 5. *Let $\mathcal{G}$ be a graph and $\mathcal{A} \subset \mathcal{G}$, with $f : \mathcal{A} \to \mathbf{R}$ such that, whenever $a, b \in \mathcal{A}$ and there is a path in $\mathcal{G}$ of length at most $m < \omega$ between a and b, $|f(a) - f(b)| \le m$. Then there is a map $f' : \mathcal{G} \to \mathbf{R}$ extending f such that, for all $a', b' \in \mathcal{G}$ such that there is a path of length at most m between a' and b', $|f'(a') - f'(b')| \le m$.*

Proof. This is a special case of the extension property of Lipschitz continuous functions: without loss of generality, we may assume $\mathcal{A}$ and $\mathcal{G}$ are connected, which makes them metric spaces if we take the distance between two vertices to be the length of the shortest path between them in $\mathcal{G}$. †

Lemma 6. *If $r : \omega \to \mathbf{R}^+$ and $u : S_r \to I$ is as above, and, for each $i < \omega$, f_i is a strictly increasing exponential mechanism with rate $k(i)$, there is $\tilde{u} : (I \times 2)^{<\omega} \to I$ extending u such that $\Lambda_{\tilde{u},f}$ is purely $(\ell, 4k(\ell)r(\ell))$-differentially private for every positive integer l.*

Proof. Combine Lemmas 3, 4, and 5. †

4 Uniform Measure Accuracy

Having established that the learners arising from exponential mechanisms are private, we now show that the logistic and Laplacian exponential mechanisms give accurate learners. All of our integrals, and therefore all of our assertions realted to accuracy, are taken with respect to the uniform Lebesgue measure on the interval.

We first note that, if ℓ is large in comparison to $r(\ell)^{-1}$, then samples of length ℓ drawn according to the uniform measure lie in S_r with high probability. As this probability depends only on $r(\ell)$ rather than on the sequence r, we state the result for the constant sequence $r(\ell) = N^{-1}$, where N is assumed to be an integer.

Lemma 7. *If $\ell, N > 0$ are integers and μ the uniform measure on I, $p_0 : (I \times 2)^{<\omega} \to I^{<\omega}$ the (post-composition with) projection map, $\mu^\ell(p_0(S_{N^{-1}})) \ge 1 - N(1 - N^{-1})^\ell$. In particular, if $\delta > 0$ and $\ell > \frac{\log \delta - \log N}{\log(1 - N^{-1})}$, then $\mu^\ell(p_0(S_{N^{-1}})) \ge 1 - \delta$.*

Proof. If the union does not equal I, there must be some k such that $\operatorname{im} \sigma \cap (kN^{-1}, (k+1)N^{-1}) = \varnothing$. For any fixed such $k < N$, $\mu^\ell\{\sigma \in I^\ell : \operatorname{im} \sigma \cap (kN^{-1}, (k+1)N^{-1}) = \varnothing\} = (1 - N^{-1})^\ell$. Consequently, the μ^ℓ-measure of their union is less than or equal to $N(1 - N^{-1})^\ell$. †

Note that, if $\sigma \in p_0 S_r$ and $A \subset I$ is upwards-closed, $(\sigma, A \circ \sigma) \in S_r$.

For $x \in I, y \in \mathbf{R}, k > 0$ we write $\mathfrak{A}_k(x, y) = \int_0^x 1 - (1 + e^{-k(z-y)})^{-1} dz + \int_x^1 (1 + e^{-k(z-y)})^{-1} dz$. This is a measure of how close the logistic function with rate k centered at y is to the characteristic function of $(x, 1]$; namely, it is the accuracy of $\Lambda_{u,f}$ when it tries to learn the set $(x, 1]$ from training data η, where f is a logistic function and $\eta \in (I \times 2)^{<\omega}$ is such that $u(\eta) = y$. To complete our argument, we must show that $\mathfrak{A}_k(x, y)$ is close to 1 when $|x - y|$ is small, which must be the case when x, y, and η are as above if $\eta \in S_r$ and $r(|\eta|)$ is small, for $|x - y| \le 2r(|\eta|)$.

An antiderivative of $(1 + e^{-k(z-y)})^{-1}$ with respect to z is $k^{-1} \log(1 + e^{k(z-y)})$. So

$$\mathfrak{A}_k(x, y) = x + k^{-1}(\log(1 + e^{k(1-y)}) + \log(1 + e^{-ky}) - 2\log(1 + e^{k(x-y)}))$$

If $x \le y$,

$$\begin{aligned} \mathfrak{A}_k(x, y) &\ge 1 - x + y - 2k^{-1} \log(1 + e^{k(x-y)}) \\ &\ge 1 - x + y - 2(\log 2)k^{-1}. \end{aligned}$$

By the symmetry of the logistic function, we have $\mathfrak{A}_k(x, y) = \mathfrak{A}_k(1 - x, 1 - y)$, and therefore whenever $x, y \in I$,

$$\mathfrak{A}_k(x, y) \ge 1 - |x - y| - 2(\log 2)k^{-1}.$$

We calculate the the accuracy $\mathfrak{B}_k(x,y)$ of $\Lambda_{u,g}(\eta)$ where g is the antiderivative of the Laplacian distribution and $u(\eta) = y$ for the set $(x,1]$ in a similar fashion, obtaining $\mathfrak{B}_k(x,y) \geq 1 - |x-y| - k^{-1}$.

Of course, $2\log 2 > 1$; the advantage of the logistic function is that it is more private in likely output regions, unlike the Laplacian, which is equally private everywhere.

Lemma 8. *For any positive integers N, ℓ, and all $\delta > 0$, if $f = (f_i)_{i<\omega}$ is a sequence of logistic or Laplacian exponential mechanisms, the rate k of f_ℓ is greater than 2, and $\ell > \frac{\log\delta - \log N}{\log(1-N^{-1})}$, then the restriction of $\Lambda_{N^{-1},f}$ (as defined in 3) to $(I\times 2)^\ell$ is a $(3(N^{-1}+2k^{-1}),\delta)$-accurate, $4kN^{-1}$-private learner for the upwards-closed subsets of I.*

Proof. We only need constrain the impact of the normalizing factor $(f(1-a)-f(-a))^{-1}$ on the accuracy. Write $f_{k,a}(x) = (1+e^{-k(x-a)})^{-1}$. For the accuracy claim, it suffices to show $3^{-1} \leq f_{k,a}(1) - f_{k,a}(0) \leq 1$ whenever $a \in I$. Note that the upper bound is obvious and, as a function of $a \in I$, $f_{k,a}(1) - f_{k,a}(0)$ attains its global minima when $a = 0$ and $a = 1$, the only other local extremum being the global maximum at $a = 2^{-1}$. So we must concern ourselves with bounding $(1+e^{-k})^{-1} - 2^{-1}$ from below, which is easy to do as the reciprocal function is convex and hence

$$(1+e^{-k})^{-1} - 2^{-1} \geq 1 - e^{-k} - 2^{-1} = 2^{-1} - e^{-k}.$$

This exceeds 3^{-1} whenever $k > 2$, and a similar computation yields the same bound for each k in the case where f_ℓ is Laplacian. †

Lemma 9. *If $\epsilon, \delta, \alpha > 0$ and $N, \ell > 2$ are positive integers such that $N > 3\epsilon^{-1}(1+8\alpha^{-1})$ and $\ell > \frac{\log\delta - \log N}{\log(1-N^{-1})}$, there is an (ϵ,δ)-accurate, $(\ell,\alpha,0)$-private learner on $(I\times 2)^\ell$ for the upwards-closed subsets of I with respect to the uniform measure on I. In particular, there is a real number C such that, for all $\alpha,\epsilon,\delta \in (0,2^{-2})$, there is an (ℓ,ϵ,δ)-accurate,(ℓ,α)-private learner on $(I\times 2)^\ell$ for some ℓ with $\ell < C\alpha^{-1}\epsilon^{-1}(\log\delta^{-1} + \log\alpha^{-1} + \log\epsilon^{-1})$.*

Proof. Put $k = 2(3^{-1}\epsilon - N^{-1})^{-1}$ in the preceding lemma. For the latter statement, put $D = 25 > 3(8+\alpha)$ and take N to be the least integer greater than $D\alpha^{-1}\epsilon^{-1} > 3\epsilon^{-1}(1+8\alpha^{-1})$. Then

$$\frac{\log\delta - \log N}{\log(1-N^{-1})} < N(\log N + \log\delta^{-1})$$

$$< (D\alpha^{-1}\epsilon^{-1}+1)(\log(D+1) + \log\alpha^{-1} + \log\epsilon^{-1} + \log\delta^{-1})$$

$$< 2D\alpha^{-1}\epsilon^{-1}(\log\alpha^{-1} + \log\epsilon^{-1} + \log\delta^{-1})$$

Therefore, there is some $\ell < 2D\alpha^{-1}\epsilon^{-1}(\log\alpha^{-1} + \log\epsilon^{-1} + \log\delta^{-1})$ such that $\Lambda_{N^{-1},f}$ with f logistic or Laplacian is $(\alpha,0)$-private and (ϵ,δ)-accurate on $(I\times 2)^\ell$. †

We can combine these into a single learner for all finite sequences in a natural way.

Lemma 10. *If C a constant as in Lemma 9 and, for each $\ell < \omega$, $h(\ell)$ is the largest integer such that $l > 3Ch(\ell)^2\log h(\ell)$, $r(\ell) = C^{-1}h(\ell)^{-2}$, $k(\ell) = 2(5^{-1}h(\ell)^{-1} - r(\ell))^{-1}$, and $f(\ell)$ is either the logistic or Laplacian exponential mechanism with rate $k(\ell)$, then $\Lambda_{\tilde{u},f}$ constructed from r and f as in Lemma 6 is $(h(\ell)^{-1}, h(\ell)^{-1})$-accurate with respect to the uniform measure and $(h(\ell)^{-1},0)$-private for all l.*

5 Nonuniform Measures

We have demonstrated a learner which is private and accurate with respect to uniform Lebesgue measure on the real unit interval I. The nonexistence of a private learner which is accurate with respect to every measure on I is implied by [3]. More precisely, the proof of Theorem 3.6 given there yields, with slight modification, a logarithmic dependency of the number of training data a learner requires to learn with a fixed accuracy and privacy on the size of a finite linear order whose upwards-closed subsets it learns.[4] We suggest that the substantive distinction between the case we have considered and the cases treated in [3] is not that one measure is uniform and the other is not, but rather that the learner is only required to be accurate with respect to one measure.

If a measure μ on the Borel subsets of I has a continuous distribution function $F(x) = \mu([0,x])$, then F is a weakly order-preserving isomorphism between the measure space I equipped with μ and the measure space arising

[4] We include the proof of our case in an appendix.

when I is equipped with the uniform measure, and the two situations are equivalent from our perspective. Composing our learners $\Lambda_{f,u}$ above with F in the appropriate fashion yields learners which are accurate with respect to μ : if $(\sigma,\tau) \in (I\times 2)^{\ell}$, then $(F\circ\sigma,\tau) \in I\times 2^{\ell}$, and if we consider the distribution function $\overline{\Theta}$ arising from integration of the probability density function Θ given in Lemma 3, $\overline{\Theta}\circ F$ is a continuous distribution function on I; if $(\overline{\Theta}\circ F)(x)$ is taken to be the likelihood of the learner selecting an infimum for the set its guesses which is less than or equal to x, one obtains a learner with the same accuracy and privacy guarantees given in the preceding two sections.

However, if the distribution function F associated to μ is not continuous (though, by the elementary theory of measures, it must be right-continuous), we have a fundamentally different learning problem. The point masses at discontinuities of F complicate Lemma 7, and the integrals appearing in the accuracy calculation must be replaced by rougher estimates. Furthermore, we must expand our hypothesis class to include all upwards-closed subsets of I, not just those that were open: with respect to the uniform measure, there is no meaningful distinction between $(a,1]$ and $[a,1]$, but, in the general case, there is.

We shall describe the outputs of our learners by specifying the likelihood $H(x)$ they accord to each $x \in I$ for belonging to the set being learned; when $\mu(x) = 0$, our learners will only output $(x,1]$, never $[x,1]$ (the choice being arbitrary), and when $\mu(x) > 0$, we accord $(x,1]$ the probability $\lim_{y\to x^+} H(y) - H(x)$ and $[x,1]$ the probability $H(x) - \lim_{y\to x^-} H(y)$. The one-sided limits exist because H is weakly increasing. Note that F is not assumed to be absolutely continuous.

Write $F_1(x) = 2^{-1}(F(x) + \lim_{y\to x^-} F(y))$ and $\mathfrak{Y} = \{(x,1]|x \in I\} \cup \{[x,1]|x \in I\}$. The inequalities $F_1 \leq F$ and, for $x < y$, $F(x) \leq F_1(y)$ are critical to the following computations.

We begin our analysis by modifying Lemma 3; the statement expresses what was written in natural language in the previous paragraph.

Lemma 11. *Suppose μ is a Borel measure on I and $F(x) = \mu([0,x])$ is its distribution function. Let $u : (I\times 2)^{<\omega} \to I$ and $d : \omega \to \mathbf{R}$ be functions such that, whenever $\eta E \nu$ in the aforementioned graph structure, $|u(\eta) - u(\nu)| \leq d(|\eta|)$, and, for each natural number i, let f_i be a strictly increasing exponential mechanism with rate $k(i)$. Then the learner $\Lambda_{u,f}$ which assigns to $\eta = (\sigma,\tau) \in (I\times 2)^i$ the measure on $\mathfrak{Y}$ corresponding to the membership likelihood function*

$$H_\eta(x) = \frac{f_i(F_1(x) - u(\eta))}{f_i(1 - u(\eta)) - f_i(-u(\eta))}$$

is purely $(i, 2k(i)d(i))$-differentially private for every i.

Proof. We must examine the measure on the two components of $\mathfrak{Y}$. Let ρ_η be the measure on I corresponding to the distribution function $\lim_{y\to x^+} H_\eta(x)$, let ν_η be the measure on I corresponding to the distribution function $\sum_{x\leq b} H(x) - \lim_{y\to x^-} H(y)$, and let ξ_η be the measure corresponding to the distribution function $\sum_{x\leq b} \lim_{y\to x^+} H(y) - H(x)$.Then put $\lambda_\eta = \rho_\eta - \nu_\eta - \xi_\eta$. It is easy to see that λ_η is a positive measure, and the measure κ on $\mathfrak{Y}$ which induces H is, for each Borel $A \subset I$,

$$\kappa_\eta\{(x,1]|x \in A\} = \lambda_\eta(A) + \xi_\eta(A)$$

$$\kappa_\eta\{[x,1]|x \in A\} = \nu_\eta(A)$$

None of these are probability measures. If, for some $\alpha > 0$, we show that λ, ξ, ν assignments are α−private, the α−privacy of κ will follow. Suppose $\eta E \theta$, $\eta = (\sigma,\tau), \theta = (\beta,\gamma)$. By Lemma 2, the denominator of H_η and the denominator of H_θ cannot differ by a factor greater than $d(i)k(i)$, so we need concern ourselves only with the numerator in each case.

Both ν_η and ξ_η concentrate on countable sets, and therefore it suffices to control the behavior of ν and ξ on single-element sets, namely those $x \in I$ where $\mu(x) > 0$. Further,

$$H_\eta(x) - \lim_{y\to x^-} H_\eta(y) = \frac{f_i(F_1(x) - u(\eta)) - \lim_{y\to x^-} f_i(F_1(y) - u(\eta))}{f_i(1 - u(\eta)) - f_i(-u(\eta))}$$

$$= \frac{\lim_{y\to x^-} \int_{F_1(y)-u(\eta)}^{F_1(x)-u(\eta)} f_i'}{f_i(1 - u(\eta)) - f_i(-u(\eta))}$$

with the analogous equation for θ, so Definition 3 yields the result. The analogous computation holds for ξ_η.

Fix some enumeration $(c_j)_{j=0}^{\infty}$ of the points x with $\mu(x) > 0$, and put $\mu^n(A) = \mu(A) - \sum_{n<j} \mu(c_j)A(c_j)$ for each Borel A. Of course, the μ_i converge uniformly to μ. Also, let F^n be the distribution function of μ^n and put $F_1^n(x) = 2^{-1}(F^n(x) + \lim_{y\to x^-} F(y))$. Then we may define

$$H_\eta^n(x) = \frac{f_i(F_1^n(x) - u(\eta))}{f_i(1 - u(\eta)) - f_i(-u(\eta))},$$

and $\lambda_\eta^n, \rho_\eta^n, \nu_\eta^n, \xi_\eta^n$ are defined for μ^n as $\lambda_\eta, \rho_\eta, \nu_\eta, \xi_\eta$ were defined for μ; all of the sequences of measures converge uniformly. Note that, for $x \in I$ and n and integer, $\lambda_\eta^n(x) = 0$. Therefore, the result will follow if we show the required inequality between λ_η^n and λ_θ^n for every i, and it suffices to check $\lambda_\theta^n(a,b] \le e^{2k(i)d(i)}\lambda_\theta^n(a,b]$ whenever $\{c_j\}_{j\le n} \cap (a,b] = \varnothing$. But, in such cases, $\lambda_\theta^n(a,b] = \rho_\theta^n(a,b]$, and the relevant inequality for ρ_θ^n and ρ_η^n is affirmed using Definition 3 much as in the case for ν_η and ν_θ. We are done. †

Suppose f_i is a logistic mechanism. Note that, if $\epsilon \in (0,1)$ and $k(i) > \epsilon^{-1}\log(\epsilon^{-1} - 1)$, $x < -\epsilon$ implies $f_i(x) < \epsilon$; if f_i is Laplacian, the implication holds whenever $k(i) > \epsilon^{-1}(-\log 2 + \log \epsilon^{-1})$.

We obtain u in a different fashion, for the S_r employed above will not quite suffice. To this end,

We apply Theorem 2 of [6] to obtain immediately

Lemma 12. *If $\ell > 0$ is an integer, $\ell > 2r(\ell)^{-2}$, and μ a Borel measure on I, $p_0 : (I\times 2)^{<\omega} \to I^{<\omega}$ the (post-composition with) projection map, and $\binom{m}{\le k}$ denotes the number of subsets of a (finite) set with m elements which contain at most k elements, then*

$$\mu^\ell(p_0(T_r^\mu)) \ge 1 - 4\binom{\ell}{\le 2}e^{-2^{-3}r(\ell)^2\ell}.$$

As in the uniform case, one has a great deal of latitude in determining how much accuracy should be sacrificed to preserve privacy; the choices we make in the theorem are reasonable asymptotically, but they require the statement be restricted to cases with at least 24^4 training data, a condition which is not forced by the fundamental structure of our method.

Lemma 13. *Write $h(x) = 2^{-1}(1 - \sqrt{1 - 24x^{-1/4}})$. For each Borel measure μ on I, there is a learner Λ which is*

$$(\ell, 3(6\ell^{-1/4}\log(h(x)^{-1} - 1) + h(x) + \ell^{-1/3}), 4\binom{\ell}{\le 2}e^{-2^{-3}\ell^{1/3}}))$$

- accurate and $(\ell, \ell^{-1/12})$−private for all upwards-closed subsets of I with respect to μ whenever $\ell > 24^4$.

Proof. Put $r(\ell) = \ell^{-1/3}$ and $k(\ell) = \ell^{1/4}$, and let $f = (f_i)_{i<\omega}$ be the sequence of logistic exponential mechanisms with rate $k(i)$. Let $T_r^\mu \subset (I\times 2)^{<\omega}$ consist of (σ,τ) with the property that $\sigma(i) \le \sigma(j) \to \tau(i) \le \tau(j)$ and, for all intervals A with $\mu A > r(|\sigma|)$, there is some i such that $\sigma(i) \in A$. We define $u : T_r^\mu \to I$ by

$$u(\sigma,\tau) = 2^{-1}(\max_{i\in\tau^{-1}(0)} F_1(\sigma(i)) + \min_{i\in\tau^{-1}(1)} F_1(\sigma(i))),$$

and the proof of Lemma 4 yields the analogous result for this u, as $F_1(b) - F_1(a) \le \mu[a,b]$; we abuse notation by referring to a fixed extension of u to all of $(I\times 2)^{<\omega}$ which maintains the Lipschitz constant also as u.[5]

Take $\Lambda_{u,f}$ as in the statement of Lemma 11; the privacy statement is the conclusion of the lemma. We must now constrain the accuracy. First we consider the measure given by only the numerator of H_η, neglecting the denominator in the statement of Lemma 11 which normalizes to a probability measure. By Lemma 12, we need concern ourselves only with the restriction of $\Lambda_{u,f}$ to T_r^μ. Suppose $\eta = (\sigma,\tau) \in T_r^\mu$ and $|\eta| = \ell$. Write $\epsilon = -f_\ell^{-1}(h(\ell))$, and let $A \subset I$ be upwards-closed. Put $a = \inf A$. Our restriction on ℓ implies $f_\ell(-2^{-1}r(\ell)) > h(\ell)$ and, hence, $r(\ell) < 2\epsilon$.

Suppose without loss of generality that $u(\eta) \le F_1(a)$. Write $B = F_1^{-1}(u(\eta) - \epsilon, u(\eta) + \epsilon)$. If $B = (s,t)$, then $\mu B = \lim_{w\to t^-} F_1(w) - \lim_{p\to s^+} F_1(p) < 2\epsilon$; if $B = [s,t]$ with $s < t$, then $\mu B = F(t) - \lim_{w\to s^-} F(w)$, and since $\lim_{p\to t^-} F(p) > u(\eta) - \epsilon$, using $F_1(t) < u(\eta) + \epsilon$, we obtain $F(t) < u(\eta) + 3\epsilon$, and the same argument goes for

[5] We changed the notation slightly; what is now r is analogous to what was before $2r$.

$\lim_{w \to s^-} F(w)$, so $\mu B < 6\epsilon$. Similar computations apply to the half-open case; what remains is the case where $B = \{s\}$. Here, either $\mu B < r(\ell)$ or s is in the image of σ. In the latter case, $|u(\eta) - F_1(s)| \geq 4^{-1}\mu(s)$, because for any $x \neq s$, we have $|F_1(x) - F_1(s)| \geq 2^{-1}\mu(s)$. (This is why we needed to use F_1 instead of F to define $\Lambda_{u,f}$.) So $\mu(s) \leq 4\epsilon$.

We have either that $\mu B < 6\epsilon$ or that $B = \{s\}$ and $\mu s < r(\ell)$. In the latter case, we clearly have an accuracy at least $1 - r(\ell) - h(\ell)$, and, in the former, we note $\mu(A^c \cap F_1^{-1}[u(\eta) + \epsilon, 1]) < r(\ell)$ because $\sigma^{-1}(A^c \cap F_1^{-1}[u(\eta) + \epsilon, 1]) = \varnothing$, so we obtain accuracy of at least $1 - h(\ell) - r(\ell) - 6\epsilon$. The factor of 3 arising in the normalization (the denominator of H_η) multiplies the loss by at most 3, and the result follows. †

References

[1] Martin Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: Association for Computing Machinery, 2016, pp. 308–318. URL: https://doi.org/10.1145/2976749.2978318.

[2] Malliaris et al. "Private PAC Learning Implies Finite Littlestone Dimension". In: *STOC Proc* (2019).

[3] A. Beimel et al. "Bounds on the sample complexity for private learning and private data release." In: *Machine Learning* 94 (2014), pp. 401–437.

[4] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Theory of Cryptography* (2006).

[5] United States Census Bureau. "DAS 2020 Redistricting Production Code". Git library of production code from the 2020 Census conducted by the USA. URL: https://github.com/uscensusbureau/DAS_2020_Redistricting_Production_Code.

[6] Chervonenkis and Vapnik. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". Trans. by Seckler. In: *Soviet Mathematics Doklady* (1971).

[7] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. FTTCS, 2014.

[8] Carey Radebaugh and Ulfar Erlingsson. "Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data". Online announcement of a TensorFlow library. URL: https://medium.com/tensorflow/introducing-tensorflow-privacy-learning-with-differential-privacy-for-training-data-b143c5e801b6.