Not all learnable distribution classes are privately learnable

Not All Learnable Distribution Classes are Privately Learnable[*]

Mark Bun[†] Gautam Kamath[‡] Argyris Mouzakis[§] Vikrant Singhal[¶]

February 6, 2024

Abstract

We give an example of a class of distributions that is learnable in total variation distance with a finite number of samples, but not learnable under (ε, δ)-differential privacy. This refutes a conjecture of Ashtiani.

1 Introduction

Given samples from a distribution $\mathcal{D}$ belonging to some class of distributions $\mathcal{H}$, can we output a distribution $\mathcal{D}'$ that is close to $\mathcal{D}$ in total variation distance? This problem, known as *distribution learning* or *density estimation*, has enjoyed significant study by a number of communities, including Computer Science, Statistics, and Information Theory (see, e.g., [DL01, KMR$^+$94, DDS12, ABDH$^+$20]).

A recent line of work studies distribution learning under *differential privacy* [DMNS06], giving sample complexity bounds for several classes of interest. However, many of these algorithms are ad hoc, exploiting idiosyncrasies of the class of interest (see, e.g., [KV18, KLSU19]). Recent efforts have succeeded in weakening assumptions and designing increasingly general learning algorithms and frameworks (see, e.g., [LKO22, KMS$^+$22b, AL22, KMV22, AAL23a]). It is natural to wonder how far this agenda can be pushed – what are the limits of private learning? Specifically, we consider the following question:

Question 1.1. Is every learnable class of distributions $\mathcal{H}$ also learnable under the constraint of (ε, δ)-differential privacy?

The answer is known to be "no" under the stronger constraint of $(\varepsilon, 0)$-DP (i.e., *pure* DP). Bun, Kamath, Steinke, and Wu [BKSW19] showed that the covering and packing numbers of a distribution class $\mathcal{H}$ give sample complexity upper and lower bounds, respectively, for learning the class $\mathcal{H}$. Consequently, this immediately gives separations between learning and $(\varepsilon, 0)$-DP learning.[1] However, they do not prove any sample complexity lower bounds for (ε, δ)-DP (i.e., *approximate* DP) learning, leaving open the possibility that every learnable distribution class is privately learnable.

On the related task of PAC learning of *functions*, a rich line of work shows that there exist strong separations between non-private learning and private learning, under both $(\varepsilon, 0)$-DP [BBKN14, FX15] and (ε, δ)-DP [BNSV15, ALMM19, BLM20]. In particular, for approximate DP, learnability

[*]Authors are listed in alphabetical order.

[†]`mbun@bu.edu`. Boston University.

[‡]`g@csail.mit.edu`. Cheriton School of Computer Science, University of Waterloo and Vector Institute.

[§]`amouzaki@uwaterloo.ca`. Cheriton School of Computer Science, University of Waterloo.

[¶]`vikrant.singhal@uwaterloo.ca`. OpenDP. Work done while a postdoc at the University of Waterloo.

[1]The simplest natural example is the class of univariate unit-variance Gaussians with unbounded mean.

is characterized by the Littlestone dimension, rather than the VC dimension as in the non-private setting. However, given substantial differences in the setting, it is unclear whether these separations have any implications for private distribution learning.

At a July 2022 workshop at the Fields Institute, Ashtiani explicitly conjectured an affirmative answer to Question 1.1: every learnable class of distributions is privately learnable [Ash22]. Indeed, as mentioned before, the community (including contributions by Ashtiani, as well as others) has designed increasingly generic algorithms for private distribution learning [AL22, TCK+22, AAL23a], often depending only on a non-private learner in a black-box manner.

We refute Ashtiani's conjecture, and give an explicit class of distributions which is learnable from a constant number of samples, but is not privately learnable with any finite number of samples.

Theorem 1.2 (Informal version of Theorem 3.8)**.** *There exists a class of distributions $\mathcal{H}$ such that, for an absolute constant c:*

1. *There exists an algorithm which, given $\mathcal{O}(1)$ samples from any distribution $\mathcal{D} \in \mathcal{H}$, outputs a $\widehat{\mathcal{D}} \in \mathcal{H}$ such that $\mathbb{P}\left[d_{\mathrm{TV}}\left(\widehat{\mathcal{D}}, \mathcal{D}\right) \leqslant c\right] \geqslant 0.9$.*
2. *Any (ε, δ)-DP mechanism that attains the same accuracy guarantee needs an infinite number of samples.*

We use a "trapdoor" construction, where the class of distributions consists of mixtures over two components. The components are entangled, in the sense that they share the same set of parameters. The first component encodes a "key" that makes it possible to identify the other component. The second component is hard to learn individually, even without privacy. In our setting, the first component will be a binary product distribution over $\{0, 1\}^d$, whereas the other component will be a distribution over $\{\pm 1, \ldots, \pm d\}$. However, we stress that d will not be fixed a-priori, in the sense that our class will include distributions where d can be any positive integer greater than 1.[2] The construction will be done in a way that the mixing weight will significantly favor the second component, but samples drawn from it will give very little information about the overall distribution. Eventually, the hardness in the private setting will be a consequence of reducing from lower bounds for private mean estimation of binary product distributions (in the appropriate error metric). We note that conceptually-similar (but technically quite different) trapdoor constructions have recently been used to show lower bounds for PAC learning [LBD23] and robust learnability [BDBKL23].

Related Work. Gaussians are often the first class studied when considering distribution learning. They have been studied under the constraint of differential privacy starting from the work of Karwa and Vadhan on estimating univariate Gaussians [KV18], with subsequent works focused on understanding the multivariate setting [KLSU19, BS19, BDKU20, LKKO21, AAK21, CWZ21, TCK+22, AL22, KMS+22b, KMV22, BKS22, KMS22a, AKT+23, HKMN23, AUZ23, KMR+23], as well as the related problem of binary product distributions [KLSU19, Sin23]. The natural generalization to learning mixtures of Gaussians has also been studied [NRS07, KSSU19, AAL21, AL22, AAL23b, AAL23a]. Some work focuses on estimating structured classes of distributions [DHS15]. Other works study broad tools for distribution learning [BKSW19, AAK21, ASZ21, TCK+22, AL22]. See [KU20] for a survey of the area.

[2] We focus on the $d \geqslant 2$ case because, for $d = 1$, the two components will have overlapping supports.

2 Preliminaries

General Notation. We denote the set of all non-zero integers by $\mathbb{Z}^*$. Additionally, given a set S, we define S^i to be the *i-fold Cartesian product of the set with itself.* We use the notation $[n] := \{1, 2, \dots, n\}$ and $[a \pm R] := [a - R, a + R]$. Also, for convenience, we will use the notations like $\left(\mathbb{R}^d\right)^n \equiv \mathbb{R}^{n \times d}$ and $\left(\{0,1\}^d\right)^n \equiv \{0,1\}^{n \times d}$. We use $\mathrm{Bern}(p)$ to denote a Bernoulli distribution with probability of success p. Furthermore, given any set S, we denote the set of all distributions over that set by $\Delta(S)$. For any distribution $\mathcal{D}$, $\mathcal{D}^{\bigotimes n}$ denotes the *product measure* where each *marginal distribution* is $\mathcal{D}$. Thus, if we are given n independent samples from $\mathcal{D}$, we write $(X_1, \dots, X_n) \sim \mathcal{D}^{\bigotimes n}$. Also, depending on the context, we may use capital Latin characters like X to denote either an individual sample from a distribution or a collection of samples $X := (X_1, \dots, X_n)$. To denote the j-th component of a vector, we will use a subscript (e.g., X_j, if the vector is X). Given a pair of distributions $\mathcal{D}_1, \mathcal{D}_2$ over a space $\mathcal{X}$, their TV-distance is defined as $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) := \sup_{A \subseteq \mathcal{X}} |\mathcal{D}_1(A) - \mathcal{D}_2(A)|$. If $\mathcal{D}_1$ and $\mathcal{D}_2$ are discrete, it holds that $d_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathcal{D}_1(x) - \mathcal{D}_2(x)|$.

We conclude this section by introducing the definition of differential privacy and its *closure under post-processing* property.

Definition 2.1 (Differential Privacy (DP) [DMNS06])**.** A mechanism $M : \mathcal{X}^n \to \mathcal{Y}$ is said to satisfy (ε, δ)-differential privacy ((ε, δ)-DP) if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ (i.e., datasets that differ in exactly one entry), we have:

$$\mathop{\mathbb{P}}_{M}[M(X) \in Y] \leqslant e^{\varepsilon} \mathop{\mathbb{P}}_{M}\left[M\left(X'\right) \in Y\right] + \delta, \quad \forall\, Y \subseteq \mathcal{Y}.$$

When $\delta = 0$, we say that M satisfies ε-differential privacy or pure differential privacy.

Lemma 2.2 (Post Processing [DMNS06])**.** *If $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-DP, and $P : \mathcal{Y} \to \mathcal{Z}$ is any randomized function, then the algorithm $P \circ M$ is (ε, δ)-DP.*

3 The Construction and Proofs

We define the class of distributions $\mathcal{H}_{w,d} := \left\{\mathcal{D}_{w,d,p} : p \in [0,1]^d\right\} \subseteq \Delta\left(\{0,1\}^d \cup \{\pm 1, \dots, \pm d\}\right)$, where each $\mathcal{D}_{w,d,p}$ has pmf $q_{w,d,p}$ with:

$$q_{w,d,p}(x) := \begin{cases} w, & \prod_{j \in [d]} p_j^{x_j} (1 - p_j)^{1 - x_j}, \forall x \in \{0,1\}^d \\ \frac{1-w}{d}, & p_1^{\frac{1+x}{2}} (1 - p_1)^{\frac{1-x}{2}}, \forall x \in \{\pm 1\} \\ \frac{1-w}{d}, & p_2^{\frac{1+\frac{x}{2}}{2}} (1 - p_2)^{\frac{1-\frac{x}{2}}{2}}, \forall x \in \{\pm 2\} \\ & \vdots \\ \frac{1-w}{d}, & p_d^{\frac{1+\frac{x}{d}}{2}} (1 - p_d)^{\frac{1-\frac{x}{d}}{2}}, \forall x \in \{\pm d\} \end{cases}.$$

Simply put, each $\mathcal{D}_{w,d,p}$ is a mixture of $d+1$ components. The first component has mixing weight w and is a binary product distribution over $\{0,1\}^d$ with probability vector p. Each of the remaining components has mixing weight $\frac{1-w}{d}$ and is a binary distribution that takes the value j with probability p_j and the value $-j$ with probability $1 - p_j$. Note, in particular, that the probability vector p is shared for both components of the distribution. In this context, the first component can be seen as the "key" to learning the distribution, because a single sample from it reveals information about the whole parameter vector, in contrast to the last d components which, taken together, play the role of the “hard distribution”, since a sample from it reveals information about only one component of the parameter vector. Our goal will be to use $\mathcal{H}_w := \bigcup_{d \geqslant 2} \mathcal{H}_{w,d}$ as the class that will lead to the separation. Specifically, we will show that the sample complexity of privately learning each $\mathcal{H}_{w,d}$ is dimension-dependent. As d grows, the sample complexity will approach infinity. At this point, we note that lower bounds shown for individual classes $\mathcal{H}_{w,d}$ are also lower bounds for $\mathcal{H}_w$ which, combined with our previous observation, implies that it's impossible to learn $\mathcal{H}_w$ with a finite number of samples.

Suppose that our target error is denoted by α. Our proof will focus on an instance of $\mathcal{H}_w$ with $w = \frac{\alpha}{2}$. Specifically, focusing on the sub-class $\mathcal{H}_{\frac{\alpha}{2},d}$ for $d \geqslant 2$, we will first show a lower bound of $\Omega\left(\frac{\sqrt{d}}{\log\left(\frac{1}{\alpha}\right)\sqrt{\alpha}\varepsilon}\right)$ for density estimation up to error α with probability of success 0.9 for this class under (ε, δ)-DP (Corollary 3.5), and then argue that the non-private sample complexity for the same task is $\mathcal{O}\left(\frac{1}{\alpha^3}\right)$ (Lemma 3.7). We conclude by formally establishing the desired separation in Theorem 3.8.

We start by showing the lower bound under privacy. Doing so involves an argument which establishes a reduction from parameter estimation for binary product distributions to density estimation for the class $\mathcal{H}_{\frac{\alpha}{2},d}$. Formulating the reduction first necessitates showing how a mechanism that performs density estimation for the class $\mathcal{H}_{\frac{\alpha}{2},d}$ can be used to construct a mechanism that estimates the parameter p of distributions in this class.

Lemma 3.1. *Let $d \geqslant 2, p \in [0,1]^d$, and $X \sim \mathcal{D}^{\otimes n}_{\frac{\alpha}{2},d,p}$. If $M \colon \left(\{0,1\}^d \cup \{\pm 1, \ldots, \pm d\}\right)^n \to \mathcal{H}_{\frac{\alpha}{2},d}$ is an (ε,δ)-DP mechanism that outputs a $\widehat{\mathcal{D}}$ such that $\mathbb{E}_{X,M}\left[d_{\mathrm{TV}}\left(\widehat{\mathcal{D}}, \mathcal{D}_{\frac{\alpha}{2},d,p}\right)\right] \leqslant \alpha \leqslant 1$, then it is possible to output a $\widehat{p} \in [0,1]^d$ such that $\mathbb{E}_{X,M}[\|\widehat{p} - p\|_1] \leqslant 2d\alpha$, while preserving (ε,δ)-DP.*

Proof. We observe that all the distributions in the class are mixtures with two components that have disjoint supports, and that the mixing weights are the same for all distributions. As a consequence, given a pair $p_1, p_2 \in [0,1]^d$, we have the following for the corresponding distributions:

$$d_{\mathrm{TV}}\left(\mathcal{D}_{\frac{\alpha}{2},d,p_1}, \mathcal{D}_{\frac{\alpha}{2},d,p_2}\right) = \frac{\alpha}{2} d_{\mathrm{TV}}\left(\bigotimes_{j\in[d]} \mathrm{Bern}(p_{1,j}), \bigotimes_{j\in[d]} \mathrm{Bern}(p_{2,j})\right) + \frac{1-\frac{\alpha}{2}}{d}\|p_1 - p_2\|_1. \tag{1}$$

Based on the above, if we have a distribution $\widehat{\mathcal{D}} \equiv \mathcal{D}_{\frac{\alpha}{2},d,\widehat{p}}$, such that $\mathbb{E}_{X,M}\left[d_{\mathrm{TV}}\left(\widehat{\mathcal{D}}, \mathcal{D}_{\frac{\alpha}{2},d,p}\right)\right] \leqslant \alpha$, it must always be the case that $\frac{1-\frac{\alpha}{2}}{d}\mathbb{E}_{X,M}[\|\widehat{p}-p\|_1] \leqslant \alpha \implies \mathbb{E}_{X,M}[\|\widehat{p}-p\|_1] \leqslant \frac{d\alpha}{1-\frac{\alpha}{2}} \leqslant 2d\alpha$. Thus, all we have to do is identify the probability vector $\widehat{p}$ that corresponds to $\widehat{\mathcal{D}}$ and output it, while privacy is preserved thanks to Lemma 2.2. ∎

To complete the reduction, we need to show how, given a mechanism that performs density estimation for the class $\mathcal{H}_{\frac{\alpha}{2},d}$, it is possible to use it to perform ℓ_1-parameter estimation for binary product distributions. This is done in the following lemma:

Lemma 3.2. *For $d \geqslant 2$, let P be a binary product distribution over $\{0,1\}^d$ with mean vector $p \in [0,1]^d$, and let $X \sim P^{\otimes n}$. If any (ε,δ)-DP mechanism $T \colon \{0,1\}^{n \times d} \to [0,1]^d$ with $\mathbb{E}_{X,T}[\|T(X) - p\|_1] \leqslant 2d\alpha$ requires at least $n \geqslant n_0$ samples, the same sample complexity lower* *bound holds for any* (ε, δ)*-DP mechanism* $M\colon \left(\{0,1\}^d \cup \{\pm 1, \dots, \pm d\}\right)^n \to \mathcal{H}_{\frac{\alpha}{2},d}$ *that satisfies* $\mathop{\mathbb{E}}_{Y,M}\left[d_{\mathrm{TV}}\left(M(Y), \mathcal{D}_{\frac{\alpha}{2},d,p}\right)\right] \leqslant \alpha \leqslant 1$*, where* $Y \sim \mathcal{D}_{\frac{\alpha}{2},d,p}^{\otimes n}$.

Proof. To establish our result, it suffices to show that estimating the parameter vector of P can be transformed into an instance of density estimation for distributions in $\mathcal{H}_{\frac{\alpha}{2},d}$, implying that lower bounds for the former problem also apply to the latter. To do so, we assume we have an (ε, δ)-DP mechanism $M\colon \left(\{0,1\}^d \cup \{\pm 1, \dots, \pm d\}\right)^n \to \mathcal{H}_{\frac{\alpha}{2},d}$ with $\mathop{\mathbb{E}}_{Y,M}\left[d_{\mathrm{TV}}\left(M(Y), \mathcal{D}_{\frac{\alpha}{2},d,p}\right)\right] \leqslant \alpha \leqslant 1$ for $Y \sim \mathcal{D}_{\frac{\alpha}{2},d,p}^{\otimes n}$. We will show how to use this mechanism to construct an (ε, δ)-DP mechanism $T\colon \{0,1\}^{n\times d} \to [0,1]^d$ with $\mathop{\mathbb{E}}_{X,T}[\|T(X) - p\|_1] \leqslant 2d\alpha$ for $X \sim P^{\otimes n}$.

The crux of the argument involves proving that, given a dataset $X \sim P^{\otimes n}$, it is possible to generate a dataset $Y \sim \mathcal{D}_{\frac{\alpha}{2},d,p}^{\otimes n}$. The mechanism T will consist of this sampling step (pre-processing), and an application of M over the resulting dataset. Appealing to Lemma 3.1 suffices to establish that T will have the desired accuracy guarantee, so the rest of the proof is devoted to describing the sampling process.

Given any datapoint X_i, we set Y_i equal to it with probability $\frac{\alpha}{2}$, or, with probability $1 - \frac{\alpha}{2}$, we choose one of the coordinates of X_i uniformly at random (say the j-th coordinate). If the j-th coordinate of X_i is equal to 1, we set $Y_i = j$. Otherwise, we set $Y_i = -j$. The resulting dataset Y will follow the desired distribution. We stress that this process preserves privacy guarantees, because changing a point of X can result in at most one point of Y changing (conditioned on the randomness involved in the conversion of X to Y). ■

At this point, we recall the following result from [KLSU19]:

Proposition 3.3. *[Lemma 6.2 from [KLSU19]] Let* p *be any vector in* $\left[\frac{1}{3}, \frac{2}{3}\right]^d$*, and let* $X := (X_1, \dots, X_n)$ *be a dataset consisting of* n *independent samples from a binary product distribution* P *over* $\{0,1\}^d$ *with mean* p*. If* $M\colon \{0,1\}^{n\times d} \to \left[\frac{1}{3}, \frac{2}{3}\right]^d$ *is an* (ε, δ)*-DP mechanism with* $\varepsilon \in [0,1]$ *and* $\delta = \mathcal{O}\left(\frac{1}{n}\right)$ *that satisfies* $\mathop{\mathbb{E}}_{X,M}\left[\|M(X) - p\|_2^2\right] \leqslant \alpha^2 \leqslant \mathcal{O}(d), \forall p \in \left[\frac{1}{3}, \frac{2}{3}\right]^d$*, it must hold that* $n \geqslant \Omega\left(\frac{d}{\alpha\varepsilon}\right)$.

While phrased in terms of mechanisms with mean-squared-error guarantees, the above result also implies a bound for ℓ_1-estimation. The connection is described in the following lemma:

Lemma 3.4. *For* $d \geqslant 2$*, an absolute constant* $C_1 > 0$*, and any* $\alpha \leqslant C_1$*, consider the class of distributions* $\mathcal{H}_{\frac{\alpha}{2},d}$*. Let* $p \in \left[\frac{1}{3}, \frac{2}{3}\right]^d$*, and let* $X \sim \mathcal{D}_{\frac{\alpha}{2},d,p}^{\otimes n}$*. If* $M\colon \left(\{0,1\}^d \cup \{\pm 1, \dots, \pm d\}\right)^n \to \mathcal{H}_{\frac{\alpha}{2},d}$ *is an* (ε, δ)*-DP mechanism with* $\varepsilon \in [0,1]$ *and* $\delta = \mathcal{O}\left(\frac{1}{n}\right)$ *that outputs a* $\widehat{\mathcal{D}}$ *such that* $\mathop{\mathbb{E}}_{X,M}\left[d_{\mathrm{TV}}\left(\widehat{\mathcal{D}}, \mathcal{D}_{\frac{\alpha}{2},d,p}\right)\right] \leqslant \alpha, \forall p \in \left[\frac{1}{3}, \frac{2}{3}\right]^d$*, it must hold that* $n \geqslant \Omega\left(\frac{\sqrt{d}}{\sqrt{\alpha}\varepsilon}\right)$.

Proof. We recall the inequality $\|x\|_2^2 \leqslant \|x\|_\infty \|x\|_1, \forall x \in \mathbb{R}^d$. This is a consequence of Hölder's inequality, but can also be shown in an elementary way by remarking that:

$$\|x\|_2^2 = \sum_{i\in[d]} x_i^2 \leqslant \max_{i\in[d]}\{|x_i|\} \sum_{i\in[d]} |x_i| = \|x\|_\infty \|x\|_1.$$

Now, let X be a dataset of size n that has been drawn i.i.d. from a binary product distribution P with mean vector p, and let $T\colon \{0,1\}^{n\times d} \to [0,1]^d$ be an (ε, δ)-DP mechanism with $\varepsilon \in [0,1], \delta = \mathcal{O}\left(\frac{1}{n}\right)$ that satisfies $\mathop{\mathbb{E}}_{X,T}[\|T(X) - p\|_1] \leqslant 2d\alpha$. We have $\|T(X) - p\|_\infty \leqslant 1$ which, by an application of the above inequality, yields $\|T(X)-p\|_2^2 \leqslant \|T(X)-p\|_1$. This implies that T satisfies the guarantee $\underset{X,T}{\mathbb{E}}\left[\|T(X)-p\|_2^2\right] \leqslant 2d\alpha$. Consequently, the lower bound of Proposition 3.3 applies to T if we set $\alpha \to \sqrt{2d\alpha}$. Then, appealing to Lemma 3.2 completes the proof. ∎

The lower bound of Lemma 3.4 also holds for mechanisms that achieve the accuracy guarantee $\underset{X,M}{\mathbb{P}}\left[d_{\mathrm{TV}}\left(\widehat{\mathcal{D}}, \mathcal{D}_{\frac{\alpha}{2},d,p}\right) \leqslant \alpha\right] \geqslant 0.9$, albeit at the cost of getting a result that's weaker by a log-factor. The argument is sketched in the proof of Theorem 6.1 of [KLSU19], so we point readers there and do not repeat it here. The resulting sample complexity bound is $n \geqslant \Omega\left(\frac{\sqrt{d}}{\log\left(\frac{1}{\alpha}\right)\sqrt{\alpha}\varepsilon}\right)$.

We summarize the above remarks in the following corollary.

Corollary 3.5. *For $d \geqslant 2$, an absolute constant $C_1 > 0$, and any $\alpha \leqslant C_1$, consider the class of distributions $\mathcal{H}_{\frac{\alpha}{2},d}$. Let $p \in \left[\frac{1}{3}, \frac{2}{3}\right]^d$, and let $X \sim \mathcal{D}_{\frac{\alpha}{2},d,p}^{\otimes n}$. If $M\colon \left(\{0,1\}^d \cup \{\pm 1, \ldots, \pm d\}\right)^n \to \mathcal{H}_{\frac{\alpha}{2},d}$ is an (ε,δ)-DP mechanism with $\varepsilon \in [0,1]$ and $\delta = \mathcal{O}\left(\frac{1}{n}\right)$ that outputs a $\widehat{\mathcal{D}}$ such that $\underset{X,M}{\mathbb{P}}\left[d_{\mathrm{TV}}\left(\widehat{\mathcal{D}}, \mathcal{D}_{\frac{\alpha}{2},d,p}\right) \leqslant \alpha\right] \geqslant 0.9, \forall p \in \left[\frac{1}{3}, \frac{2}{3}\right]^d$, it must hold that $n \geqslant \Omega\left(\frac{\sqrt{d}}{\log\left(\frac{1}{\alpha}\right)\sqrt{\alpha}\varepsilon}\right)$.*

Remark 3.6. While the lower bounds in the above statements are phrased in terms of *proper learners*, they also imply the same bounds against *improper learners*. If computation is not a concern, an improper learner can be converted to a proper one by enumerating over all densities in the class and projecting to whichever one is closest with respect to the TV-distance. Since the TV-distance satisfies the triangle inequality, this can lead to the error increasing by a factor of 2.

We now proceed to argue that the non-private sample complexity of proper density estimation with respect to the TV-distance for the class $\mathcal{H}_{\frac{\alpha}{2},d}$ is independent of d.

Lemma 3.7. *For $d \geqslant 2$, there exists an algorithm $\mathcal{A}\colon \left(\{0,1\}^d \cup \{\pm 1, \ldots, \pm d\}\right)^n \to \mathcal{H}_{\frac{\alpha}{2},d}$ which, given a dataset $X \sim \mathcal{D}_{\frac{\alpha}{2},d,p}^{\otimes n}$ of size $n = \mathcal{O}\left(\frac{\log\left(\frac{1}{\beta}\right)}{\alpha^3}\right)$, outputs a distribution $\widehat{\mathcal{D}} \equiv \mathcal{D}_{\frac{\alpha}{2},d,\widehat{p}} \in \mathcal{H}_{\frac{\alpha}{2},d}$ such that:*

$$\underset{X}{\mathbb{P}}\left[d_{\mathrm{TV}}\left(\mathcal{D}_{\frac{\alpha}{2},d,\widehat{p}}, \mathcal{D}_{\frac{\alpha}{2},d,p}\right) \leqslant \alpha\right] \geqslant 1-\beta.$$

Proof. By (1), we have:

$$d_{\mathrm{TV}}\left(\mathcal{D}_{\frac{\alpha}{2},d,\widehat{p}}, \mathcal{D}_{\frac{\alpha}{2},d,p}\right) = \frac{\alpha}{2} d_{\mathrm{TV}}\left(\bigotimes_{j\in[d]} \mathrm{Bern}(\widehat{p}_j), \bigotimes_{j\in[d]} \mathrm{Bern}(p_j)\right) + \frac{1-\frac{\alpha}{2}}{d}\|\widehat{p}-p\|_1.$$

Based on the above, in order to attain error α in TV-distance, it suffices to (1) estimate $\bigotimes_{j\in[d]} \mathrm{Bern}(p_j)$ up to error 1 in TV-distance, and (2) estimate the vector p up to error $\frac{d\alpha}{2}$ in ℓ_1-distance. Statement (1) holds trivially, since all distributions are at TV-distance 1 from each other, so we focus on (2).

For (2), it holds that $\|\widehat{p}-p\|_1 \leqslant \sqrt{d}\|\widehat{p}-p\|_2$, so it suffices to have a $\widehat{p}$ such that $\|\widehat{p}-p\|_2 \leqslant \frac{\sqrt{d}\alpha}{2}$. Assume, now, that we are given m samples drawn i.i.d. from a binary product distribution, and that we want to estimate its parameter vector within ℓ_2-error α with probability at least $1-\frac{\beta}{2}$. It is a folklore fact that $m = \Theta\left(\frac{d+\log\left(\frac{1}{\beta}\right)}{\alpha^2}\right)$ samples, are both necessary and sufficient for this task, with the bound being attained by taking the sample mean. Thus, setting $\alpha \to \frac{\sqrt{d}\alpha}{2}$ yields $\Theta\left(\frac{d+\log\left(\frac{1}{\beta}\right)}{d\alpha^2}\right)$, which is dominated by $\mathcal{O}\left(\frac{\log\left(\frac{1}{\beta}\right)}{\alpha^2}\right)$. Consequently, in order to get $\|\widehat{p} - p\|_2 \leqslant \frac{\sqrt{d}\alpha}{2}$ in our setting, it suffices to have $m = \mathcal{O}\left(\frac{\log\left(\frac{1}{\beta}\right)}{\alpha^2}\right)$ samples from the first component (the binary product distribution). For that reason, assume that, for each datapoint X_i we draw from $\mathcal{D}_{\frac{\alpha}{2},d,p}$, we have an associated random variable $Z_i \sim \text{Bern}\left(\frac{\alpha}{2}\right)$ which becomes 1 if X_i comes from the first component. We assume now that we have n samples with $\frac{n\alpha}{2} \geqslant m$. We will show that $n = \mathcal{O}\left(\frac{\log\left(\frac{1}{\beta}\right)}{\alpha^3}\right)$ suffices to ensure that the event $\sum_{i\in[n]} Z_i < m$ doesn't happen, except with probability at most $\frac{\beta}{2}$. The Hoeffding bound implies that:

$$\mathbb{P}\left[\sum_{i\in[n]} Z_i < m\right] \leqslant \mathbb{P}\left[\left|\sum_{i\in[n]} Z_i - \frac{n\alpha}{2}\right| \geqslant \frac{n\alpha}{2} - m\right] \leqslant e^{-\frac{(n\alpha-2m)^2}{2n}}.$$

To ensure that the above is upper-bounded by $\frac{\beta}{2}$, it suffices to have $n \geqslant \frac{2\left(2\alpha m+\log\left(\frac{2}{\beta}\right)\right)}{\alpha^2} = \mathcal{O}\left(\frac{\log\left(\frac{1}{\beta}\right)}{\alpha^3}\right)$. By a union bound, the total probability of failure is upper-bounded by β, completing the proof. ■

We are now ready to establish our main result.

Theorem 3.8. *Given any $\mathcal{D} \in \mathcal{H}_{\frac{C_1}{2}}$, we have:*

1. *There exists an algorithm $\mathcal{A}\colon \bigcup_{d\geqslant 2}\left(\{0,1\}^d \cup \{\pm 1,\ldots,\pm d\}\right)^n \to \mathcal{H}_{\frac{C_1}{2}}$ which, given $n = \mathcal{O}(1)$ samples drawn i.i.d. from $\mathcal{D}$, outputs a $\widehat{\mathcal{D}} \in \mathcal{H}_{\frac{C_1}{2}}$ such that $\mathbb{P}_X\left[d_{\text{TV}}\left(\widehat{\mathcal{D}},\mathcal{D}\right) \leqslant \frac{C_1}{2}\right] \geqslant 0.9$.*
2. *For finite n, there exists no (ε,δ)-DP mechanism $M\colon \bigcup_{d\geqslant 2}\left(\{0,1\}^d \cup \{\pm 1,\ldots,\pm d\}\right)^n \to \mathcal{H}_{\frac{C_1}{2}}$ with $\varepsilon \in [0,1], \delta = \mathcal{O}\left(\frac{1}{n}\right)$ which, given a dataset $X \sim \mathcal{D}^{\otimes n}$, outputs a $\widehat{\mathcal{D}} \in \mathcal{H}_{\frac{C_1}{2}}$ that satisfies $\mathbb{P}_{X,M}\left[d_{\text{TV}}\left(\widehat{\mathcal{D}},\mathcal{D}\right) \leqslant \frac{C_1}{2}\right] \geqslant 0.9$.*

Proof. Let a (potentially adversarially chosen) $\mathcal{D} \equiv \mathcal{D}_{\frac{C_1}{2},d,p} \in \mathcal{H}_{\frac{C_1}{2},d}$ be our ground truth.

Without privacy constraints, all the algorithm $\mathcal{A}$ has to do is look at the samples to identify the number of components d, and then calculate the sample mean corresponding to the samples from the first component (as we did in the proof of Lemma 3.7). The desired guarantee is immediate by the guarantees of that lemma.

Under privacy, we will establish our result by contradiction. Assume that, for some finite n, there exists an (ε,δ)-DP mechanism $M\colon \bigcup_{d\geqslant 2}\left(\{0,1\}^d \cup \{\pm 1,\ldots,\pm d\}\right)^n \to \mathcal{H}_{\frac{C_1}{2}}$ with $\varepsilon \in [0,1], \delta = \mathcal{O}\left(\frac{1}{n}\right)$ which, given $X \sim \mathcal{D}^{\otimes n}$, outputs a $\widehat{\mathcal{D}} \in \mathcal{H}_{\frac{C_1}{2}}$ such that $\mathbb{P}_{X,M}\left[d_{\text{TV}}\left(\widehat{\mathcal{D}},\mathcal{D}\right) \leqslant \frac{C_1}{2}\right] \geqslant 0.9$. We note that $\widehat{\mathcal{D}}$ might not be in $\mathcal{H}_{\frac{C_1}{2},d}$ (since the output range is assumed to be the entire $\mathcal{H}_{\frac{C_1}{2}}$). However, working as we described in Remark 3.6, we can round the output to an element of $\mathcal{H}_{C_{\frac{1}{2}},d}$, with the TV-distance between the resulting distribution and the ground truth now being C_1 (the privacy guarantee is preserved thanks to Lemma 2.2). Then, by Corollary 3.5, it must be the case that $n \geqslant \Omega\left(\frac{\sqrt{d}}{\varepsilon}\right)$. This must hold for every $d \geqslant 2$, so taking $d \to \infty$ leads to a contradiction. ■

4 Acknowledgments

We would like to thank Thomas Steinke for helpful discussions related to the reduction from ℓ_2-estimation to ℓ_1-estimation, as well as the reviewers at ALT for their comments which led to the improvement of the paper's presentation. We would further like to thank the Fields Institute for Research in Mathematical Sciences, and the organizers of the Summer 2022 Workshop on Differential Privacy and Statistical Data Analysis, where Ashtiani presented this conjecture.

MB is supported by NSF CNS-2046425 and a Sloan Research Fellowship. GK, AM, and VS are supported by a Canada CIFAR AI Chair, an NSERC Discovery Grant, and an unrestricted gift from Google. AM is further supported by a scholarship from the Onassis Foundation (Scholarship ID: F ZT 053-1/2023-2024).

References

[AAK21] Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 185–216. JMLR, Inc., 2021. 2

[AAL21] Ishaq Aden-Ali, Hassan Ashtiani, and Christopher Liaw. Privately learning mixtures of axis-aligned gaussians. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021. 2

[AAL23a] Mohammad Afzali, Hassan Ashtiani, and Christopher Liaw. Mixtures of gaussians are privately learnable with a polynomial number of samples. *arXiv preprint arXiv:2309.03847*, 2023. 1 and 2

[AAL23b] Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded gaussian mixture models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pages 1018–1040. JMLR, Inc., 2023. 2

[ABDH+20] Hassan Ashtiani, Shai Ben-David, Nicholas JA Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *Journal of the ACM*, 67(6):32:1–32:42, 2020. 1

[AKT+23] Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a Gaussian: Efficient, robust and optimal. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing*, STOC '23, New York, NY, USA, 2023. ACM. 2

[AL22] Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning Gaussians and beyond. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1075–1076, 2022. 1 and 2

[ALMM19] Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *Proceedings of the 51st Annual ACM Symposium on the Theory of Computing*, STOC '19, pages 852–860, New York, NY, USA, 2019. ACM. 1

[Ash22] Hassan Ashtiani. Private learning of gaussians and their mixtures. https://www.youtube.com/watch?v=bmNjm0lx50I, July 2022. 2

[ASZ21] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 48–78. JMLR, Inc., 2021. 2

[AUZ23] Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. *arXiv preprint arXiv:2302.01855*, 2023. 2

[BBKN14] Amos Beimel, Hai Brenner, Shiva Prasad Kasiviswanathan, and Kobbi Nissim. Bounds on the sample complexity for private learning and private data release. *Machine Learning*, 94(3):401–437, 2014. 1

[BDBKL23] Shai Ben-David, Alex Bie, Gautam Kamath, and Tosca Lechner. Distribution learnability and robustness. In *Advances in Neural Information Processing Systems 36*, NeurIPS '23. Curran Associates, Inc., 2023. 2

[BDKU20] Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 14475–14485. Curran Associates, Inc., 2020. 2

[BKS22] Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022. 2

[BKSW19] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 156–167. Curran Associates, Inc., 2019. 1 and 2

[BLM20] Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '20, pages 389–402, Washington, DC, USA, 2020. IEEE Computer Society. 1

[BNSV15] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society. 1

[BS19] Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 181–191. Curran Associates, Inc., 2019. 2

[CWZ21] T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021. 2

[DDS12] Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A. Servedio. Learning Poisson binomial distributions. In *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing*, STOC '12, pages 709–728, New York, NY, USA, 2012. ACM. 1

[DHS15] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2566–2574. Curran Associates, Inc., 2015. 2

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer, 2001. 1

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer. 1 and 3

[FX15] Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. *SIAM Journal on Computing*, 44(6):1740–1764, 2015. 1

[HKMN23] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on the Theory of Computing*, STOC '23, New York, NY, USA, 2023. ACM. 2

[KLSU19] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1853–1902, 2019. 1, 2, 5, and 6

[KMR+94] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the 26th Annual ACM Symposium on the Theory of Computing*, STOC '94, pages 273–282, New York, NY, USA, 1994. ACM. 1

[KMR+23] Gautam Kamath, Argyris Mouzakis, Matthew Regehr, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A bias-variance-privacy trilemma for statistical estimation. *arXiv preprint arXiv:2301.13334*, 2023. 2

[KMS22a] Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022. 2

[KMS+22b] Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 544–572, 2022. 1 and 2

[KMV22] Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 723–777, 2022. 1 and 2

[KSSU19] Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman. Differentially private algorithms for learning mixtures of separated Gaussians. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 168–180. Curran Associates, Inc., 2019. 2

[KU20] Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020. 2

[KV18] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. 1 and 2

[LBD23] Tosca Lechner and Shai Ben-David. Impossibility of characterizing distribution learning–a simple solution to a long-standing problem. *arXiv preprint arXiv:2304.08712*, 2023. 2

[LKKO21] Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021. 2

[LKO22] Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1167–1246, 2022. 1

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, STOC '07, pages 75–84, New York, NY, USA, 2007. ACM. 2

[Sin23] Vikrant Singhal. A polynomial time, pure differentially private estimator for binary product distributions. *arXiv preprint arXiv:2304.06787*, 2023. 2

[TCK+22] Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22, pages 21828–21863. JMLR, Inc., 2022. 2