Private Hypothesis Selection

Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu

***Abstract*—We provide a differentially private algorithm for hypothesis selection. Given samples from an unknown probability distribution P and a set of m probability distributions $\mathcal{H}$, the goal is to output, in a ε-differentially private manner, a distribution from $\mathcal{H}$ whose total variation distance to P is comparable to that of the best such distribution (which we denote by α). The sample complexity of our basic algorithm is $O\left(\frac{\log m}{\alpha^2} + \frac{\log m}{\alpha\varepsilon}\right)$, representing a minimal cost for privacy when compared to the non-private algorithm. We also can handle infinite hypothesis classes $\mathcal{H}$ by relaxing to (ε, δ)-differential privacy.**

We apply our hypothesis selection algorithm to give learning algorithms for a number of natural distribution classes, including Gaussians, product distributions, sums of independent random variables, piecewise polynomials, and mixture classes. Our hypothesis selection procedure allows us to generically convert a cover for a class to a learning algorithm, complementing known learning lower bounds which are in terms of the size of the packing number of the class. As the covering and packing numbers are often closely related, for constant α, our algorithms achieve the optimal sample complexity for many classes of interest. Finally, we describe an application to private distribution-free PAC learning.

***Index Terms*—differential privacy, hypothesis selection, density estimation**

I. Introduction

We consider the problem of *hypothesis selection*: given samples from an unknown probability distribution, select a distribution from some fixed set of candidates which is "close" to the unknown distribution in some appropriate distance measure. Such situations can arise naturally in a number of settings. For instance, we may have a number of different methods which work under various circumstances, which are not known in advance. One option is to run all the methods to generate a set of hypotheses, and pick the best from this set afterwards. Relatedly, an algorithm may branch its behavior based on a number of "guesses," which will similarly result in a set of candidates, corresponding to the output at the end of each branch. Finally, if we know that the underlying distribution belongs to some (parametric) class, it is possible to essentially enumerate the class (also known as a *cover*) to create a collection of hypotheses. Observe that this last example is quite general, and this approach can give generic learning algorithms for many settings of interest.

This problem of hypothesis selection has been extensively studied (see, e.g., [1], [2], [3], [4]), resulting in algorithms with a sample complexity which is *logarithmic* in the number of hypotheses. Such a mild dependence is critical, as it facilitates sample-efficient algorithms even when the number of candidates may be large. These initial works have triggered a great deal of study into hypothesis selection with additional considerations, including computational efficiency, understanding the optimal approximation factor, adversarial robustness, and weakening access to the hypotheses (e.g., [5], [6], [7], [8], [9], [10], [11], [12]).

However, in modern settings of data analysis, data may contain sensitive information about individuals. Some examples of such data include medical records, GPS location data, or private message transcripts. As such, we would like to perform statistical inference in these settings without revealing significant information about any particular individual's data. To this end, there have been many proposed notions of data privacy, but perhaps the gold standard is that of *differential privacy* [13]. Informally, differential privacy requires that, if a single datapoint in the dataset is changed, then the distribution over outputs produced by the algorithm should be similar (see Definition II.4). Differential privacy has seen widespread adoption, including deployment by Apple [14], Google [15], and the US Census Bureau [16].

This naturally raises the question of whether one can perform hypothesis selection under the constraint of differential privacy, while maintaining a logarithmic dependence on the size of the cover. Such a tool would allow us to generically obtain private learning results for

Authors are in alphabetical order. A preliminary version of this paper appeared in Advances in Neural Information Processing Systems 32 (NeurIPS 2019). This work was supported in part by Google and Microsoft Research Fellowships as part of the Simons-Berkeley Research Fellowship program, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and in part while visiting Microsoft Research, Redmond, and the Simons Institute for the Theory of Computing at UC Berkeley.

M. Bun was with the Simons Institute for the Theory of Computing, and is now with Boston University. mbun@bu.edu.

G. Kamath was with the Simons Institute for the Theory of Computing, and is now with the University of Waterloo. gckamath@uwaterloo.ca.

T. Steinke was with IBM, and is now with Google. phs@thomas-steinke.net.

Z. S. Wu was with the University of Minnesota, and is now with Carnegie Mellon University. zstevenwu@cmu.edu.

Manuscript received April 19, 2005; revised August 26, 2015.

Copyright (c) 2017 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

a wide variety of settings.

A. Results

Our main results answer this in the affirmative: we provide differentially private algorithms for selecting a good hypothesis from a set of distributions. The output distribution is competitive with the best distribution, and the sample complexity is bounded by the logarithm of the size of the set. The following is a basic version of our main result.

Theorem I.1. *Let $\mathcal{H} = \{H_1, \ldots, H_m\}$ be a set of probability distributions. Let $D = \{X_1, \ldots, X_n\}$ be a set of samples drawn independently from an unknown probability distribution P. There exists an ε-differentially private algorithm (with respect to the dataset D) which has following guarantees. Suppose there exists a distribution $H^* \in \mathcal{H}$ such that $d_{\mathrm{TV}}(P, H^*) \leq \alpha$. If $n = \Omega\left(\frac{\log m}{\alpha^2} + \frac{\log m}{\alpha\varepsilon}\right)$, then the algorithm will output a distribution $\hat{H} \in \mathcal{H}$ such that $d_{\mathrm{TV}}(P, \hat{H}) \leq (3+\zeta)\alpha$ with probability at least $9/10$, for any constant $\zeta > 0$. The running time of the algorithm is $O(nm^2)$.*

The sample complexity of this problem without privacy constraints is $O\left(\frac{\log m}{\alpha^2}\right)$, and thus the additional cost for ε-differential privacy is an additive $O\left(\frac{\log m}{\alpha\varepsilon}\right)$. We consider this cost to be minimal; in particular, the dependence on m is unchanged. Note that the running time of our algorithm is $O(nm^2)$ – we conjecture it may be possible to reduce this to $\tilde{O}(nm)$ as has been done in the non-private setting [7], [8], [9], [11], though we have not attempted to perform this optimization. Regardless, our main focus is on the sample complexity rather than the running time, since any method for generic hypothesis selection requires $\Omega(m)$ time, thus precluding efficient algorithms when m is large. Note that the approximation factor of $(3+\zeta)\alpha$ is effectively tight. That is, even in the infinite sample limit and without the constraint of privacy, information theoretically, one can not achieve a better approximation than 3α [4], [5].[1] Theorem I.1 requires prior knowledge of the value of α, though we can use this to obtain an algorithm with similar guarantees which does not (Theorem III.5).

It is possible to improve the guarantees of this algorithm in two ways (Theorem IV.1). First, if the distributions are nicely structured, the former term in the sample complexity can be reduced from $O(\log m/\alpha^2)$ to $O(d/\alpha^2)$, where d is a VC-dimension-based measure of the complexity of the collection of distributions. Second, if there are few hypotheses which are close to the true distribution, then we can pay only logarithmically in this number, as opposed to the total number of hypotheses. These modifications allow us to handle instances where m may be very large (or even infinite), albeit at the cost of weakening to approximate differential privacy to perform the second refinement. A technical discussion of our methods is in Section I-B, our basic approach is covered in Section III, and the version with all the bells and whistles appears in Section IV.

From Theorem I.1, we immediately obtain Corollary I.2 which applies when $\mathcal{H}$ itself may not be finite, but admits a finite cover with respect to total variation distance.

Corollary I.2. *Suppose there exists an α-cover $\mathcal{C}_\alpha$ of a set of distributions $\mathcal{H}$, and that we are given a set of samples $X_1, \ldots, X_n \sim P$, where $d_{\mathrm{TV}}(P, \mathcal{H}) \leq \alpha$. For any constant $\zeta > 0$, there exists an ε-differentially private algorithm (with respect to the input $\{X_1, \ldots, X_n\}$) which outputs a distribution $H^* \in \mathcal{C}_\alpha$ such that $d_{\mathrm{TV}}(P, H^*) \leq (6 + 2\zeta)\alpha$ with probability $\geq 9/10$, as long as*

$$n = \Omega\left(\frac{\log|\mathcal{C}_\alpha|}{\alpha^2} + \frac{\log|\mathcal{C}_\alpha|}{\alpha\varepsilon}\right).$$

Informally, this says that if a hypothesis class has an α-cover $\mathcal{C}_\alpha$, then there is a private learning algorithm for the class which requires $O(\log|\mathcal{C}_\alpha|)$ samples. Note that our algorithm works even if the unknown distribution is only *close* to the hypothesis class. This is useful when we may have model misspecification, or when we require adversarial robustness. (We also give an extension of this algorithm which gives guarantees in the *semi-agnostic* learning model; see Section III-D for details.) The requirements for this theorem to apply are minimal, and thus it generically provides learning algorithms for a wide variety of hypothesis classes. That said, in non-private settings, the sample complexity given by this method is rather lossy: as an extreme example, there is no finite-size cover of univariate Gaussian distributions with unbounded parameters, so this approach does not give a finite-sample algorithm. That said, it is well-known that $O(1/\alpha^2)$ samples suffice to estimate a Gaussian in total variation distance. In the private setting, our theorem incurs a cost which is somewhat necessary: in particular, it is folklore that any pure ε-differentially private learning algorithm must pay a cost which is logarithmic in the packing number of the class (for completeness, see Lemma V.1). Due to the relationship between packing and covering numbers (Lemma V.2), this implies that up to a constant factor relaxation in the learning accuracy, our results are tight (Theorem V.3). Further discussion appears in Sections V.

Given Corollary I.2, in Section VI, we derive new learning results for a number of classes. Our main applications are for d-dimensional Gaussian and product distributions. Informally, we obtain $\tilde{O}(d)$ sample algorithms

[1] Note that this can be brought down to 2α if one instead outputs a *mixture* of $H_i \in \mathcal{H}$ [12].

for learning a product distribution and a Gaussian with known covariance (Corollaries VI.3 and VI.10), and an $\tilde{O}(d^2)$ algorithm for learning a Gaussian with unknown covariance (Corollary VI.11). These improve on recent results by Kamath, Li, Singhal, and Ullman [17] in two different ways. First, as mentioned before, our results are semi-agnostic, so we can handle when the distribution is only *close* to a product or Gaussian distribution. Second, our results hold for pure $(\varepsilon, 0)$-differential privacy, which is a stronger notion than ε^2-zCDP as considered in [17]. In this weaker model, they also obtained $\tilde{O}(d)$ and $\tilde{O}(d^2)$ sample algorithms, but the natural modifications to achieve ε-DP incur extra $\mathrm{poly}(d)$ factors.[2] [17] also showed $\tilde{\Omega}(d)$ lower bounds for Gaussian and product distribution estimation in the even weaker model of (ε, δ)-differential privacy. Thus, our results show that the dimension dependence for these problems is unchanged for essentially any notion of differential privacy. In particular, our results show a previously-unknown separation between mean estimation of product distributions and non-product distributions under pure $(\varepsilon, 0)$-differential privacy; see Remark VI.4.

We also apply Theorem IV.1 to obtain algorithms for learning Gaussians under (ε, δ)-differential privacy, with no bounds on the mean and variance parameters. More specifically, we provide algorithms for learning multivariate Gaussians with unknown mean and known covariance (Corollary VI.13), and univariate Gaussians with both unknown mean and variance (Corollary VI.15). For the former problem, we manage to avoid dependences which arise due to the application of advanced composition (similar to Remark VI.4).

To demonstrate the flexibility of our approach, we also give private learning algorithms for sums of independent random variables (Corollaries VI.20 and VI.22) and piecewise polynomials (Corollary VI.29). To the best of our knowledge, the former class of distributions has not been considered in the private setting, and we rely on covering theorems from the non-private literature. Private learning algorithms for the latter class, piecewise polynomials, have been studied by Diakonikolas, Hardt, and Schmidt [18]. They provide sample and time efficient algorithms for histogram distributions (i.e., piecewise constant distributions), and claim similar results for general piecewise polynomials. Their method depends heavily on rather sophisticated algorithms for the non-private version of this problem [19]. In constrast, we can obtain comparable sample complexity bounds from just the existence of a cover and elementary VC dimension arguments, which we derive in a fairly self-contained manner.

We additionally give algorithms for learning mixtures of any coverable class (Corollary VI.32). In particular, this immediately implies algorithms for learning mixtures of Gaussians, product distributions, and all other classes mentioned above.

To conclude our applications, we discuss a connection to PAC learning (Corollary VI.34). It is known that the sample complexity of differentially private distribution-free PAC learning can be higher than that of non-private learning. However, this gap does not exist for distribution-specific learning, where the learning algorithm knows the distribution of (unlabeled) examples, as both sample complexities are characterized by VC dimension. Private hypothesis selection allows us to address an intermediate situation where the distribution of unlabeled examples is not known exactly, but is known to come (approximately) from a class of distributions. When this class has a small cover, we are able to recover sample complexity guarantees for private PAC learning which are comparable to the non-private case.

B. Techniques

Non-privately, most algorithms for hypothesis selection involve a tournament-style approach. We conduct a number of pairwise comparisons between distributions, which may either have a winner and a loser, or may be declared a draw. Intuitively, a distribution will be declared the winner of a comparison if it is much closer than the alternative to the unknown distribution, and a tie will be declared if the two distributions are comparably close. The algorithm will output any distribution which never loses a comparison. A single comparison between a pair of hypotheses requires $O(1/\alpha^2)$ samples, and a Chernoff plus union bound argument over the $O(m^2)$ possible comparisons increases the sample complexity to $O(\log m/\alpha^2)$. In fact, we can use uniform convergence arguments to reduce this sample complexity to $O(d/\alpha^2)$, where d is the VC dimension of the $2\binom{m}{2}$ sets (the "Scheffé" sets) defined by the subsets of the domain where the PDF of one distribution dominates another. Crucially, we must reuse the same set of samples for all comparisons to avoid paying polynomially in the number of hypotheses.

A private algorithm for this problem requires additional care. Since a single comparison is based on the number of samples which fall into a particular subset of the domain, the sensitivity of the underlying statistic is low, and thus privacy may seem easily achievable at first glance. However, the challenge comes from the fact that the same samples are reused for all pairwise comparisons, thus greatly increasing the sensitivity: changing a single datapoint could flip the result of every comparison! In order to avoid this pitfall, we instead

[2] Roughly, this is due to the fact that the Laplace and Gaussian mechanism are based on ℓ_1 and ℓ_2 sensitivity, respectively, and that there is a $\sqrt{d}$-factor relationship between these two norms, in the worst case.

carefully construct a score function for each hypothesis, namely, the minimum number of points that must be changed to cause the distribution to lose any comparison. For this to be a useful score function, we must show that the best hypothesis will win all of its comparisons by a large margin. We can then use the Exponential Mechanism [20] to select a distribution with high score.

Further improvements can be made if we are guaranteed that the number of "good" hypotheses (i.e., those that have total variation distance from the true distribution bounded by $(3+\zeta)\alpha$) is at most some parameter k, and if we are willing to relax to approximate differential privacy. The parameter k here is related to the doubling dimension of the hypothesis class with respect to total variation distance. If we randomly assign the hypotheses to $\Omega(k^2)$ buckets, with high probability, no bucket will contain more than one good hypothesis. We can identify a bucket containing a good hypothesis using a similar method based on the exponential mechanism as described above. Moreover, since we are likely to only have one "good" hypothesis in the chosen bucket, this implies a significant gap between the best and second-best scores in that bucket. This allows us to use stability-based techniques [21], [22], and in particular the GAP-MAX algorithm of Bun, Dwork, Rothblum, and Steinke [23], to identify an accurate distribution.

C. Related Work

Our main result builds on a long line of work on non-private hypothesis selection. One starting point for the particular style of approach we consider here is [1], which was expanded on in [2], [3], [4]. Since then, there has been study into hypothesis selection under additional considerations, including computational efficiency, understanding the optimal approximation factor, adversarial robustness, and weakening access to the hypotheses [5], [6], [7], [8], [9], [10], [11], [12]. Our private algorithm examines the same type of problem, with the additional constraint of differential privacy.

Perhaps the most closely related work is that of Canonne, Kamath, McMillan, Smith, and Ullman [24], which focuses on the case of private simple hypothesis testing. This is a more restricted setting than we consider in this paper, as it focuses on the case where we are trying to decide between $m = 2$ hypotheses, and we are guaranteed that the unknown distribution is one of these two hypotheses. However, in this setting, they are able to get an instance-by-instance characterization of the sample complexity, depending on both the total variation and Hellinger distance between the two distributions.

There has recently been a great deal of interest in differentially private distribution learning. In the central model, most relevant are [18], which gives algorithms for learning structured univariate distributions, and [25], [17], which focus on learning Gaussians and binary product distributions. [26] also studies private statistical parameter estimation. Privately learning mixtures of Gaussians was considered in [27], [28]. The latter paper (which is concurrent with the present work) gives a computationally efficient algorithm for the problem, but with a worse sample complexity, and incomparable accuracy guarantees (they require a separation condition, and perform clustering and parameter estimation, while we do proper learning). [29] give an algorithm for learning distributions in Kolmogorov distance. Upper and lower bounds for learning the mean of a product distribution over the hypercube in ℓ_∞-distance include [30], [31], [13], [32]. [33] focuses on estimating properties of a distribution, rather than the distribution itself. [34] gives an algorithm which allows one to estimate asymptotically normal statistics with optimal convergence rates, but no finite sample complexity guarantees. There has also been a great deal of work on distribution learning in the local model of differential privacy [35], [36], [37], [38], [39], [40], [41], [42]. For further coverage of differentially private statistics, see [43].

Non-privately, there has been a significant amount of work on learning specific classes of distributions. The PAC-style formulation of the problem we consider originated in [44]. While learning Gaussians and product distributions can be considered folklore at this point, some of the other classes we learn have enjoyed more recent study. For instance, learning sums of independent random variables was recently considered in [6] toward the problem of learning Poisson Binomial Distributions (PBDs). Since then, there has been additional work on learning PBDs and various generalizations [45], [46], [47], [48], [49], [50].

Piecewise polynomials are a highly-expressive class of distributions, and they can be used to approximate a number of other univariate distribution classes, including distributions which are multi-modal, concave, convex, log-concave, monotone hazard rate, Gaussian, Poisson, Binomial, and more. Algorithms for learning such classes are considered in a number of papers, including [51], [52], [53], [54], [19].

There has also been a great deal of work on learning mixtures of distribution classes, particularly mixtures of Gaussians. There are many ways the objective of such a problem can be defined, including clustering [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], parameter estimation [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], proper learning [79], [80], [7], [8], [10], [81], and improper learning [52]. Our work falls into the line on proper learning: the algorithm is given a set of samples from a mixture of Gaussians, and must output a mixture of Gaussians which is close in total variation distance.

1) Subsequent Work: Since the initial appearance of this work, there have been several relevant results in the surrounding area. Most pertinent is the work of Aden-Ali, Ashtiani, and Kamath [82], which improves our main hypothesis selection result by improving the constant approximation factor, making the algorithm agnostic, and having a simpler analysis, albeit at the cost of increasing the running time from quadratic to cubic in the number of hypotheses. The authors also argues that, given a small cover for each distribution in the space, there exists a cover which is locally-small everywhere as required by Theorem IV.1, thus allowing them to learn unbounded Gaussians with arbitrary covariances.

Our results and techniques have seen use for other problems as well. Kamath, Singhal, and Ullman [83] use a similar approach based on pairwise comparisons in order to perform mean estimation, rather than our work which focuses on density estimation. Liu et al. [84] use our main algorithm to estimate discrete distributions in a type of federated learning setting.

Hypothesis selection has since been studied in other privacy models. Informally speaking, our work shows that the sample complexity of hypothesis selection under central differential privacy is $O(\log m)$. Gopi et al. [85] study the problem under the stronger notion of local differential privacy [86], [87], [88], showing that the sample complexity is $\tilde{\Theta}(m)$, an exponential increase in cost. Under various notions of pan-privacy [89] and the shuffled model [90], [91], which are intermediate to local and central differential privacy, the sample complexity of m-wise simple hypothesis testing (the easier version of hypothesis selection, where the unknown distribution is equal to one of the given distributions) was shown to be $\tilde{\Theta}(\sqrt{m})$ [92].

D. Organization

We begin in Section II with preliminaries. In Section III, we give a basic algorithm for private hypothesis selection, via the exponential mechanism. In Section IV, we extend this approach in two ways: by using VC dimension arguments to reduce the sample complexity for sets of hypotheses with additional structure, and combining this with a GAP-MAX algorithm to achieve non-trivial guarantees for infinite hypothesis classes. Section V shows that our approach leads to algorithms which essentially match lower bounds for most distribution classes (in the constant α regime). We consider applications in Section VI: through a combination of arguments about covers and VC dimension, we derive algorithms for learning a number of classes of distributions, as well as describe an application to private PAC learning. Finally, we conclude in Section VII with open questions.

II. Preliminaries

We start with some preliminaries and definitions.

Definition II.1. *The* total variation distance *or* statistical distance *between P and Q is defined as*

$$\begin{aligned} d_{\mathrm{TV}}(P,Q) &= \max_{S\subseteq\Omega} P(S)-Q(S) \\ &= \frac{1}{2}\int_{x\in\Omega} |P(x)-Q(x)|dx \\ &= \frac{1}{2}\|P-Q\|_1 \in [0,1]. \end{aligned}$$

Moreover, if $\mathcal{H}$ is a set of distributions over a common domain, we define $d_{\mathrm{TV}}(P,\mathcal{H}) = \inf_{H\in\mathcal{H}} d_{\mathrm{TV}}(P,H)$.

Throughout this paper, we consider packings and coverings of sets of distributions with respect to total variation distance.

Definition II.2. *A γ-*cover *of a set of distributions $\mathcal{H}$ is a set of distributions $\mathcal{C}_\gamma$, such that for every $H\in\mathcal{H}$, there exists some $P\in\mathcal{C}_\gamma$ such that $d_{\mathrm{TV}}(P,H)\le\gamma$.*

*A γ-*packing *of a set of distributions $\mathcal{H}$ is a set of distributions $\mathcal{P}_\gamma\subseteq\mathcal{H}$, such that for every pair of distributions $P,Q\in\mathcal{P}_\gamma$, we have that $d_{\mathrm{TV}}(P,Q)>\gamma$.*

In this paper, we present semi-agnostic learning algorithms.

Definition II.3. *An algorithm is said to be an α-*semi-agnostic learner *for a class $\mathcal{H}$ if it has the following guarantees. Suppose we are given $X_1,\ldots,X_n\sim P$, where $d_{\mathrm{TV}}(P,\mathcal{H})\le \mathrm{OPT}$. The algorithm must output some distribution $\hat{H}$ such that $d_{\mathrm{TV}}(P,H)\le c\cdot\mathrm{OPT}+O(\alpha)$, for some constant $c\ge 1$. If $c=1$, then the algorithm is said to be agnostic.*

Now we define differential privacy. We say that D and D' are neighboring datasets, denoted $D\sim D'$, if D and D' differ by at most one observation. Informally, differential privacy requires that the algorithm has close output distributions when run on any pair of neighboring datasets. More formally:

Definition II.4 ([13])**.** *A randomized algorithm $T: X^*\to\mathcal{R}$ is* (ε,δ)-differentially private *if for all $n\ge 1$, for all neighboring datasets $D,D'\in X^n$, and for all events $S\subseteq\mathcal{R}$,*

$$\Pr\left[T(D)\in S\right] \le e^{\varepsilon}\Pr[T(D')\in S]+\delta\,.$$

If $\delta=0$, we say that T is ε-differentially private.

We will also use the related notion of concentrated differential privacy:

Definition II.5 ([93], [94])**.** *A randomized algorithm $T: X^*\to\mathcal{R}$ satisfies* ρ-zero-concentrated differential privacy *if for all* $n \geq 1$, *for all neighboring datasets* $D, D' \in X^n$, *and for all* $\alpha \in (1, \infty)$,

$$R_\alpha(M(D)||M(D')) \leq \rho\alpha,$$

where $R_\alpha(M(D)||M(D'))$ *is the* α*-Rényi divergence between* $M(D)$ *and* $M(D')$.[3]

The exponential mechanism [20] is a powerful ε-differentially private mechanism for selecting an approximately best outcome from a set of alternatives, where the quality of an outcome is measured by a score function relating each alternative to the underlying dataset. Letting $\mathcal{R}$ be the set of possible outcomes, a score function $q : X^* \times \mathcal{R} \to \mathbb{R}$ maps each pair consisting of a dataset and an outcome to a real-valued score. The exponential mechanism $\mathcal{M}_E$ instantiated with a dataset D, a score function q, and a privacy parameter ε selects an outcome r in $\mathcal{R}$ with probability proportional to $\exp\left(\varepsilon q(D, r)/(2\Delta(q))\right)$, where $\Delta(q)$ is the sensitivity of the score function defined as

$$\Delta(q) = \max_{r \in \mathcal{R}, D \sim D'} |q(D, r) - q(D', r)| \,.$$

Theorem II.6 ([20])**.** *For any input dataset* D*, score function* q *and privacy parameter* $\varepsilon > 0$*, the exponential mechanism* $\mathcal{M}_E(D, q, \varepsilon)$ *is* ε*-differentially private, and with probability at least* $1 - \beta$*, selects an outcome* $r \in \mathcal{R}$ *such that*

$$q(D, r) \geq \max_{r' \in \mathcal{R}} q(D, r') - \frac{2\Delta(q)\log(|\mathcal{R}|/\beta)}{\varepsilon}.$$

III. A First Method for Private Hypothesis Selection

In this section, we present our first algorithm for private hypothesis selection and obtain the following result.

Theorem I.1. *Let* $\mathcal{H} = \{H_1, \dots, H_m\}$ *be a set of probability distributions. Let* $D = \{X_1, \dots, X_n\}$ *be a set of samples drawn independently from an unknown probability distribution* P*. There exists an* ε*-differentially private algorithm (with respect to the dataset* D*) which has following guarantees. Suppose there exists a distribution* $H^* \in \mathcal{H}$ *such that* $d_{\mathrm{TV}}(P, H^*) \leq \alpha$*. If* $n = \Omega\left(\frac{\log m}{\alpha^2} + \frac{\log m}{\alpha\varepsilon}\right)$*, then the algorithm will output a distribution* $\hat{H} \in \mathcal{H}$ *such that* $d_{\mathrm{TV}}(P, \hat{H}) \leq (3 + \zeta)\alpha$ *with probability at least* $9/10$*, for any constant* $\zeta > 0$*. The running time of the algorithm is* $O(nm^2)$*.*

Note that the sample complexity bound above scales logarithmically with the size of the hypothesis class. In Section IV, we will provide a stronger result (which subsumes the present one as a special case) that can handle certain infinite hypothesis classes. For sake of exposition, we begin in this section with the basic algorithm.

A. Pairwise Comparisons

We first present a subroutine which compares two hypothesis distributions. This subroutine is due to Daskalakis, Diakonikolas, and Servedio [6], and is essentially a modification of previous methods (e.g., [4]) to allow for draws. Let H and H' be two distributions over domain $\mathcal{X}$ and consider the following set, which is called the *Scheffé set*:

$$\mathcal{W}_1 = \{x \in \mathcal{X} \mid H(x) > H'(x)\}$$

Define $p_1 = H(\mathcal{W}_1)$, $p_2 = H'(\mathcal{W}_1)$, and $\tau = P(\mathcal{W}_1)$ to be the probability masses that H, H', and P place on $\mathcal{W}_1$, respectively. It follows that $p_1 > p_2$ and $p_1 - p_2 = d_{\mathrm{TV}}(H, H')$.[4]

Algorithm 1: PAIRWISE CONTEST: PC$(H, H', D, \zeta, \alpha)$

```
Input: Two hypotheses $H$ and $H'$, input dataset
  $D$ of size $n$ drawn i.i.d. from target distribution
  $P$, approximation parameter $\zeta > 0$, and
  accuracy parameter $\alpha \in (0, 1)$.
Initialize: Compute the fraction of points that
  fall into $\mathcal{W}_1$: $\hat{\tau} = \frac{1}{n}|\{x \in D \mid x \in \mathcal{W}_1\}|$.
If $p_1 - p_2 \leq (2 + \zeta)\alpha$, return "Draw".
Else If $\hat{\tau} > p_1 - (1 + \zeta/2)\alpha$, return $H$ as the
  winner.
Else If $\hat{\tau} < p_2 + (1 + \zeta/2)\alpha$, return $H'$ as the
  winner.
Else return "Draw".
```

Now consider the function $\Gamma_\zeta(H, H', D)$ of this ordered pair of hypotheses, which is defined to be n if $p_1 - p_2 \leq (2+\zeta)\alpha$, and $n \cdot \max\{0, \hat{\tau} - (p_2 + (1+\zeta/2)\alpha)\}$ otherwise. When the two hypotheses are sufficiently far apart (i.e., $d_{\mathrm{TV}}(H, H') > (2 + \zeta)\alpha$), $\Gamma_\zeta(H, H', D)$ is essentially the number of points one needs to change in D to make H' the winner.

Lemma III.1. *Let* P, H, H' *be distributions as above. With probability at least* $1 - 2\exp(-n\zeta^2\alpha^2/8)$ *over the random draws of* D *from* P^n*,* $\hat{\tau}$ *satisfies* $|\hat{\tau} - \tau| < \zeta\alpha/4$*, and if* $d_{\mathrm{TV}}(P, H) \leq \alpha$*, then* $\Gamma_\zeta(H, H', D) > \zeta\alpha n/4$*.*

Proof. By applying Hoeffding's inequality, we know that with probability at least $1 - 2\exp(-n\zeta^2\alpha^2/8)$, $|\tau - \hat{\tau}| <$

[3] Given two probability distributions P, Q over Ω, $R_\alpha(P||Q) = \frac{1}{\alpha-1}\log\left(\sum_{x\in\Omega} P(x)^\alpha Q(x)^{1-\alpha}\right)$.

[4] For simplicity of our exposition, we will assume that we can evaluate the two quantities p_1 and p_2 exactly. In general, we can estimate these quantities to arbitrary accuracy, as long as, for each hypothesis H, we can evaluate the density of each point under H and also draw samples from H.

$\zeta\alpha/4$. We condition on this event for the remainder of the proof. Consider the following two cases. In the first case, suppose that $p_1 - p_2 \leq (2+\zeta)\alpha$. Then we know that $\Gamma_\zeta(H, H', D) = n > \alpha n$. In the second case, suppose that $p_1 - p_2 > (2+\zeta)\alpha$. Since $d_{\mathrm{TV}}(P, H) \leq \alpha$, we know that $|p_1 - \tau| \leq \alpha$, and so $|p_1 - \hat{\tau}| < (1 + \zeta/4)\alpha$. Since $p_1 > p_2 + (2+\zeta)\alpha$, we also have $\hat{\tau} > p_2 + (1+3\zeta/4)\alpha$. It follows that $\Gamma_\zeta(H, H', D) = n(\hat{\tau} - (p_2 + (1+\zeta/2)\alpha)) > \zeta\alpha n/4$. □

B. Naïve Approach via Laplace Mechanism

We first sketch a naïve approach for private hypothesis selection, based on the primitive in Algorithm 1. This is a privatization of a similar approach which appeared in [7], though the idea behind the approach is older, e.g., [4] – the [7] approach differs slightly since it employs a comparison procedure which allows ties, as we do. Later, Lemma III.6 describes the approach and privatization of [4] in more detail, which are morally equivalent to what we discuss here.

A non-private algorithm for selection from m hypotheses would run Algorithm 1 on each pair of hypotheses, either outputting a winner between the two distributions, or declaring a tie in the case when the total variation distance between the two distributions is small. The algorithm would output any distribution which never loses a comparison. Correctness of this algorithm relies on the empirical masses in all $O(m^2)$ Scheffé sets being estimated up to an additive $O(\alpha)$, which, by Hoeffding's inequality, happens with constant probability when $n \geq O\left(\frac{\log m}{\alpha^2}\right)$. Crucially, we reuse the same set of samples for all comparisons. With this in hand, it is not hard to show that a distribution H which is α-close to P will never lose a comparison, and any distribution H' which is $c\alpha$-far from P (for an appropriately chosen constant $c > 1$) will lose its comparison with H, thus ensuring that the winning distribution will be $c\alpha$-close to P.

Now, we consider how to privatize this algorithm. Each of the $O(m^2)$ comparisons is based on the quantity $\hat{\tau} = \frac{1}{n}|\{x \in D \mid x \in \mathcal{W}\}|$, where $\mathcal{W}$ is the Scheffé set between the two distributions H and H'. To make a single comparison ε-differentially private, we would have to add Laplace noise of order $O\left(\frac{1}{\varepsilon n}\right)$ to this quantity. However, since we reuse the same set of samples for all comparisons, in order to make the result of all $O(m^2)$ comparisons ε-differentially private, the basic composition property of differential privacy would prescribe adding Laplace noise of order $O\left(\frac{m^2}{\varepsilon n}\right)$ to the quantity used in each comparison. To bound the noise error of all comparisons simultaneously by $O(\alpha)$, we thus require $n \geq O\left(\frac{m^2 \log m}{\alpha\varepsilon}\right)$, and the rest of the analysis is then identical to before.

A formalization of this argument allows us to arrive at the following theorem. The accuracy bound is of the appropriate form, but the cost of privacy is an exponential increase in the sample complexity.

Theorem III.2. *Let $\mathcal{H} = \{H_1, \ldots, H_m\}$ be a set of probability distributions. Let $D = \{X_1, \ldots, X_n\}$ be a set of samples drawn independently from an unknown probability distribution P. There exists an ε-differentially private algorithm (with respect to the dataset D) which has following guarantees. Suppose there exists a distribution $H^* \in \mathcal{H}$ such that $d_{\mathrm{TV}}(P, H^*) \leq \alpha$. If $n = \Omega\left(\frac{\log m}{\alpha^2} + \frac{m^2 \log m}{\alpha\varepsilon}\right)$, then the algorithm will output a distribution $\hat{H} \in \mathcal{H}$ such that $d_{\mathrm{TV}}(P, \hat{H}) \leq O(\alpha)$ with probability at least $9/10$. The running time of the algorithm is $O(nm^2)$.*

C. Selection via Exponential Mechanism

In light of the definition of the pairwise comparison defined above, we consider the following score function $S\colon \mathcal{H} \times \mathcal{X}^n$, such that for any $H_j \in \mathcal{H}$ and dataset D,

$$S(H_j, D) = \min_{H_k \in \mathcal{H}} \Gamma_\zeta(H_j, H_k, D). \tag{1}$$

Roughly speaking, $S(H_j, D)$ is the minimum number of points required to change in D in order for H_j to lose at least one pairwise contest against a different hypothesis. When the hypothesis H_j is very close to every other distribution, such that all pairwise contests return "Draw," then the score will be n.

```
Algorithm 2: PRIVATE HYPOTHESIS SELECTION: PHS($\mathcal{H}, D, \varepsilon$)
Input: Dataset $D$, a collection of hypotheses
  $\mathcal{H} = \{H_1, \ldots, H_m\}$, privacy parameter $\varepsilon$.
Output a random hypothesis $\hat{H} \in \mathcal{H}$ such that for
  each $H_j$
      $\Pr[\hat{H} = H_j] \propto \exp\left(\frac{S(H_j, D)}{2\varepsilon}\right)$
  where $S(H_j, D)$ is defined in (1).
```

Lemma III.3 (Privacy)**.** *For any $\varepsilon > 0$ and collection of hypotheses $\mathcal{H}$, the algorithm PHS$(\mathcal{H}, \cdot, \varepsilon)$ satisfies ε-differential privacy.*

Proof. First, observe that for any pairs of hypotheses H_j, H_k, $\Gamma_\zeta(H_j, H_k, \cdot)$ has sensitivity 1. As a result, the score function S is also 1-sensitive. Then the result directly follows from the privacy guarantee of the exponential mechanism (Theorem II.6). □

Lemma III.4 (Utility)**.** *Fix any $\alpha, \beta \in (0, 1)$, and $\zeta > 0$. Suppose that there exists $H^* \in \mathcal{H}$ such* *that* $d_{\mathrm{TV}}(P, H^*) \leq \alpha$. *Then with probability* $1-\beta$ *over the sample* D *and the algorithm PHS, we have that PHS*$(\mathcal{H}, D)$ *outputs an hypothesis* $\hat{H}$ *such that* $d_{\mathrm{TV}}(P, \hat{H}) \leq (3+\zeta)\alpha$*, as long as the sample size satisfies*

$$n \geq \frac{8\ln(4m/\beta)}{\zeta^2\alpha^2} + \frac{8\ln(2m/\beta)}{\zeta\alpha\varepsilon}.$$

Proof. First, consider the m pairwise contests between H^* and every candidate in $\mathcal{H}$. Let $\mathcal{W}_j = \{x \in \mathcal{X} \mid H^*(x) > H_j(x)\}$ be the collection of Scheffé sets. For any event $W \subseteq \mathcal{X}$, let $\hat{P}(W)$ denote the empirical probability of event W on the dataset D. By Lemma III.1 and an application of the union bound, we know that with probability at least $1 - 2m\exp(-n\zeta^2\alpha^2/8)$ over the draws of D, $|P(\mathcal{W}_j) - \hat{P}(\mathcal{W}_j)| \leq \zeta\alpha/4$ and $\Gamma_\zeta(H^*, H_j, D) > \zeta\alpha n/4$ for all $H_j \in \mathcal{H}$. In particular, the latter event implies that $S(H^*, D) > \zeta\alpha n/4$.

Next, by the utility guarantee of the exponential mechanism (Theorem II.6), we know that with probability at least $1-\beta/2$, the output hypothesis satisfies

$$\begin{aligned} S(\hat{H}, D) &\geq S(H^*, D) - \frac{2\ln(2m/\beta)}{\varepsilon} \\ &> \zeta\alpha n/4 - \frac{2\ln(2m/\beta)}{\varepsilon}. \end{aligned}$$

Then as long as $n \geq \frac{8\ln(4m/\beta)}{\zeta^2\alpha^2} + \frac{8\ln(2m/\beta)}{\zeta\alpha\varepsilon}$, we know that with probability at least $1-\beta$, $S(\hat{H}, D) > 0$. Let us condition on this event, which implies that $\Gamma_\zeta(\hat{H}, H^*, D) > 0$. We will now show that $d_{\mathrm{TV}}(\hat{H}, H^*) \leq (2+\zeta)\alpha$, which directly implies that $d_{\mathrm{TV}}(\hat{H}, P) \leq (3+\zeta)\alpha$ by the triangle inequality. Suppose to the contrary that $d_{\mathrm{TV}}(\hat{H}, H^*) > (2+\zeta)\alpha$. Then by the definition of Γ_ζ, $\hat{P}(\hat{\mathcal{W}}) > H^*(\hat{\mathcal{W}}) + (1+\zeta/2)\alpha$, where $\hat{\mathcal{W}} = \{x \in \mathcal{X} \mid \hat{H}(x) > H^*(x)\}$. Since $|P(\hat{\mathcal{W}}) - \hat{P}(\hat{\mathcal{W}})| \leq \zeta\alpha/4$, we have $P(\hat{\mathcal{W}}) > H^*(\hat{\mathcal{W}}) + (1+\zeta/4)\alpha$, which is a contradiction to the assumption that $d_{\mathrm{TV}}(P, H^*) \leq \alpha$. □

D. Obtaining a Semi-Agnostic Algorithm

Theorem I.1 shows that given a hypothesis class $\mathcal{H}$ and samples from an unknown distribution P, we can privately find a distribution $\hat{H} \in \mathcal{H}$ with $d_{\mathrm{TV}}(P, \hat{H}) \leq (3+\zeta)\alpha$ *provided* that we know $d_{\mathrm{TV}}(P, \mathcal{H}) \leq \alpha$. But what if we are not promised that P is itself close to $\mathcal{H}$? We would like to design a private hypothesis selection algorithm for the more general semi-agnostic setting, where for any value of $\mathrm{OPT} := d_{\mathrm{TV}}(P, \mathcal{H})$, we are able to privately identify a distribution $\hat{H} \in \mathcal{H}$ with $d_{\mathrm{TV}}(P, \hat{H}) \leq c \cdot \mathrm{OPT} + \alpha$ for some universal constant c. Our goal will be to do this with sample complexity which is still logarithmic in $|\mathcal{H}|$.

Our strategy for handling this more general setting is by a reduction to that of Theorem I.1. We run that algorithm $T = O(\log(1/\alpha))$ times, doubling the choice of α in each run and producing a sequence of candidate hypotheses $H_1, \ldots, H_T$. By the guarantees of Theorem I.1, there is some candidate H_t with $d_{\mathrm{TV}}(P, H_t) \leq 2(3+\zeta)\,\mathrm{OPT}$. The remaining task is to approximately select the best candidate from $H_1, \ldots, H_T$. This is done by implementing a private version of the Scheffé tournament which is itself semi-agnostic, but has a very poor (quadratic) dependence on the number of candidates T.

We prove the following result, which gives a semi-agnostic learner whose sample complexity is comparable to that of Theorem I.1.

Theorem III.5. *Let* $\alpha, \beta, \varepsilon \in (0,1)$*, and* $\zeta > 0$ *be a constant. Let* $\mathcal{H}$ *be a set of* m *distributions and let* P *be a distribution with* $d_{\mathrm{TV}}(P, \mathcal{H}) = \mathrm{OPT}$*. There is an* ε*-differentially private algorithm which takes as input* n *samples from* P *and with probability at least* $1-\beta$*, outputs a distribution* $\hat{H} \in \mathcal{H}$ *with* $d_{\mathrm{TV}}(P, \hat{H}) \leq 18(3+\zeta)\,\mathrm{OPT} + \alpha$*, as long as* $n \geq O\left(\frac{\log(m/\beta) + \log\log(1/\alpha)}{\alpha^2} + \frac{\log m + \log^2(1/\alpha)\cdot(\log(1/\beta) + \log\log(1/\alpha))}{\alpha\varepsilon}\right)$*. The running time of the algorithm is* $O(m^2 n \log(1/\alpha) + n\log^2(1/\alpha))$.

As discussed above, the algorithm relies on the following variant with a much worse dependence on m.

Lemma III.6. *Let* $\alpha, \beta, \varepsilon \in (0,1)$*. There is an* ε*-differentially private algorithm which takes as input* n *samples from* P *and with probability at least* $1-\beta$*, outputs a distribution* $\hat{H} \in \mathcal{H}$ *with* $d_{\mathrm{TV}}(P, \hat{H}) \leq 9\,\mathrm{OPT} + \alpha$*, as long as*

$$n \geq O\left(\frac{\log(m/\beta)}{\alpha^2} + \frac{m^2\log(m/\beta)}{\alpha\varepsilon}\right).$$

The running time of the algorithm is $O(m^2 n)$.

Proof sketch.. We use a different variation of the Scheffé tournament which appears in [4]. Non-privately, the algorithm works as follows. For every pair of hypotheses $H, H' \in \mathcal{H}$ with Scheffé set $\mathcal{W}_{H,H'} = \{x \in \mathcal{X} \mid H(x) > H'(x)\}$, let $H(\mathcal{W}_{H,H'})$, $H'(\mathcal{W}_{H,H'})$, and $P(\mathcal{W}_{H,H'})$ denote the probability masses of H, H', P on $\mathcal{W}_{H,H'}$, respectively. Moreover, let $\hat{P}(\mathcal{W}_{H,H'})$ denote the fraction of points in the input sample D which lie in $\mathcal{W}_{H,H'}$. We declare H to be the winner of the pairwise contest between H and H' if $|H(\mathcal{W}_{H,H'}) - \hat{P}(\mathcal{W}_{H,H'})| < |H'(\mathcal{W}_{H,H'}) - \hat{P}(\mathcal{W}_{H,H'})|$. Otherwise, we declare H' to be the winner. The algorithm outputs the hypothesis $\hat{H}$ which wins the most pairwise contests (breaking ties arbitrarily).

To make this algorithm ε-differentially private, we replace $\hat{P}(\mathcal{W}_{H,H'})$ in each pairwise contest with the $(\varepsilon/\binom{m}{2})$-differentially private estimate $c_{H,H'} =$

$\hat{P}(\mathcal{W}_{H,H'}) + \mathrm{Lap}(\binom{m}{2}/\varepsilon n)$. By the composition guarantees of differential privacy, the algorithm as a whole is ε-differentially private.

The analysis of Devroye and Lugosi [4, Theorem 6.2] shows that the (private) Scheffé tournament outputs a hypothesis $\hat{H}$ with

$$d_{\mathrm{TV}}(\hat{H}, P) \le 9\,\mathrm{OPT} + 16 \max_{H,H' \in \mathcal{H}} |P(\mathcal{W}_{H,H'}) - c_{H,H'}|\,.$$

Fix an arbitrary pair H, H'. A Chernoff bound shows that $|P(\mathcal{W}_{H,H'}) - \hat{P}(\mathcal{W}_{H,H'})| \le \alpha/32$ with probability at least $1 - \beta/(2m^2)$ as long as $n \ge O(\ln(m/\beta)/\alpha^2)$. Moreover, properties of the Laplace distribution guarantee $|c_{H,H'} - \hat{P}(\mathcal{W}_{H,H'})| \le \alpha/32$ with probability at least $1 - \beta/(2m^2)$ as long as $n \ge O(m^2 \log(m/\beta)/\alpha\varepsilon)$. The triangle inequality and a union bound over all pairs H, H' complete the proof. □

Proof of Theorem III.5. We now combine the private hypothesis selection algorithm of Theorem I.1 with the expensive semi-agnostic learner of Lemma III.6 to prove Theorem III.5. Define sequences $\alpha_1 = \alpha/126, \alpha_2 = 2\alpha/126, \dots, \alpha_T = 2^{T-1}\alpha/126$ and $\varepsilon_1 = \varepsilon/4, \varepsilon_2 = \varepsilon/8, \dots, \varepsilon_T = 2^{-(T+1)}\varepsilon$ for $T = \lceil \log_2(1/\alpha) \rceil + 1$. For each $t = 1, \dots, T$, let H_t denote the outcome of a run of Algorithm 2 using accuracy parameter α_t and privacy parameter ε_t. Finally, use the algorithm of Lemma III.6 to select a hypothesis from $H_0, \dots, H_T$ using accuracy parameter α and privacy parameter $\varepsilon/2$.

Privacy of this algorithm follows immediately from composition of differential privacy. We now analyze its sample complexity guarantee. By Lemma III.4, we have that all T runs of Algorithm 2 succeed simultaneously with probability at least $1 - \beta/2$ as long as $n \ge O\left(\frac{\log(m/\beta)+\log\log(1/\alpha)}{\alpha^2} + \frac{\log(m/\beta)+\log\log(1/\alpha)}{\alpha\varepsilon}\right)$. Condition on this event occurring. Recall that success of run t of Algorithm 2 means that if $\mathrm{OPT} \in (\alpha_{t-1}, \alpha_t]$, then $d_{\mathrm{TV}}(P, H_t) \le (3+\zeta)\alpha_t \le 2(3+\zeta)\,\mathrm{OPT}$. Meanwhile, if $\mathrm{OPT} \le \alpha_1 = \alpha/126$, then we have $d_{\mathrm{TV}}(P, H_1) \le \alpha/18$. Hence, regardless of the value of OPT, there exists a run t such that $d_{\mathrm{TV}}(P, H_t) \le 2(3+\zeta)\,\mathrm{OPT} + \alpha/18$. The algorithm of Lemma III.6 is now, with probability at least $1 - \beta/2$, able to select a hypothesis $\hat{H}$ with $d_{\mathrm{TV}}(P, \hat{H}) \le 9 d_{\mathrm{TV}}(P, H_t) + \alpha/2 \le 18(3+\zeta)\,\mathrm{OPT} + \alpha$ as long as $n \ge O\left(\frac{\log(1/\beta)+\log\log(1/\alpha)}{\alpha^2} + \frac{\log^2(1/\alpha)\cdot(\log(1/\beta)+\log\log(1/\alpha))}{\alpha\varepsilon}\right)$. This gives the asserted sample complexity guarantee. □

IV. An Advanced Method for Private Hypothesis Selection

In Section III, we provided a simple algorithm whose sample complexity grows logarithmically in the size of the hypothesis class. We now demonstate that this dependence can be improved and, indeed, we can handle infinite hypothesis classes given that their VC dimension is finite and that the cover has small doubling dimension.

To obtain this improved dependence on the hypothesis class size, we must make two improvements to the analysis and algorithm. First, rather than applying a union bound over all the pairwise contests to analyse the tournament, we use a uniform convergence bound in terms of the VC dimension of the Scheffé sets. Second, rather than use the exponential mechanism to select a hypothesis, we use a "GAP-MAX" algorithm [23]. This takes advantage of the fact that, in many cases, even for infinite hypothesis classes, only a handful of hypotheses will have high scores. The GAP-MAX algorithm need only pay for the hypotheses that are close to optimal. To exploit this, we must move to a relaxation of pure differential privacy which is not subject to strong packing lower bounds (as we describe in Section V). Specifically, we consider approximate differential privacy, although results with an improved dependence are also possible under various variants of concentrated differential privacy [93], [94], [95], [23].

Theorem IV.1. *Let $\mathcal{H}$ be a set of probability distributions on $\mathcal{X}$. Let d be the VC dimension of the set of functions $f_{H,H'} : \mathcal{X} \to \{0,1\}$ defined by $f_{H,H'}(x) = 1 \iff H(x) > H'(x)$ where $H, H' \in \mathcal{H}$. There exists a (ε, δ)-differentially private algorithm which has following guarantee. Let $D = \{X_1, \dots, X_n\}$ be a set of private samples drawn independently from an unknown probability distribution P. Let $k = |\{H \in \mathcal{H} : d_{\mathrm{TV}}(H, P) \le 7\alpha\}|$. Suppose there exists a distribution $H^* \in \mathcal{H}$ such that $d_{\mathrm{TV}}(P, H^*) \le \alpha$. If $n = \Omega\left(\frac{d+\log(1/\beta)}{\alpha^2} + \frac{\log(k/\beta)+\min\{\log|\mathcal{H}|,\log(1/\delta)\}}{\alpha\varepsilon}\right)$, then the algorithm will output a distribution $\hat{H} \in \mathcal{H}$ such that $d_{\mathrm{TV}}(P, \hat{H}) \le 7\alpha$ with probability at least $1 - \beta$.*

Alternatively, we can demand that the algorithm be $\frac{1}{2}\varepsilon^2$-concentrated differentially private if $n = \Omega\left(\frac{d+\log(1/\beta)}{\alpha^2} + \frac{\log(k/\beta)+\sqrt{\log|\mathcal{H}|}}{\alpha\varepsilon}\right)$.

Comparing Theorem IV.1 to Theorem I.1, we see that the first (non-private) $\log|\mathcal{H}|$ term is replaced by the VC dimension d and the second (private) $\log|\mathcal{H}|$ term is replaced by $\log k + \log(1/\delta)$. Here k is a measure of the "local" size of the hypothesis class $\mathcal{H}$; its definition is similar to that of the doubling dimension of the hypothesis class under total variation distance.

We note that the $\log(1/\delta)$ term could be large, as the privacy failure probability δ should be cryptographically small. Thus our result includes statements for pure differential privacy (by using the other term in the minimum with $\delta = 0$) and also concentrated differential privacy. Note that, since d and $\log k$ can be upper-bounded by $O(\log|\mathcal{H}|)$, this result supercedes the guarantees of

Theorem I.1.

A. VC Dimension

We begin by reviewing the definition of Vapnik-Chervonenkis (VC) dimension and its properties.

Definition IV.2 (VC dimension [96]). *Let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \to \{0,1\}$. The VC dimension of $\mathcal{F}$ is defined to be the largest d such that there exist $x_1, \cdots, x_d \in \mathcal{X}$ and $f_1, \cdots, f_{2^d} \in \mathcal{H}$ such that for all $1 \leq i < j \leq 2^d$ there exists $1 \leq k \leq d$ such that $f_i(x_k) \neq f_j(x_k)$.*

For our setting, we must extend the definition of VC dimension from function families to hypothesis classes.

Definition IV.3 (VC dimension of hypothesis class). *Let $\mathcal{H}$ be a set of probability distributions on a space $\mathcal{X}$. For $H, H' \in \mathcal{H}$, define $f_{H,H'} : \mathcal{X} \to \{0,1\}$ by $f(x) = 1 \iff H(x) > H'(x)$. Define $\mathcal{F}(\mathcal{H}) = \{f_{H,H'} : H, H' \in \mathcal{H}\}$. We define the VC dimension of $\mathcal{H}$ to be the VC dimension of $\mathcal{F}(\mathcal{H})$.*[5]

The key property of VC dimension is the following uniform convergence bound, which we use in place of a union bound.

Theorem IV.4 (Uniform Convergence [97]). *Let $\mathcal{F}$ be a set of functions $f : \mathcal{X} \to \{0,1\}$ with VC dimension d. Let P be a distribution on $\mathcal{X}$. Then*

$$\Pr_{D \leftarrow P^n}\left[\sup_{f \in \mathcal{F}} |f(D) - f(P)| \leq \alpha\right] \geq 1 - \beta$$

whenever $n = \Omega\left(\frac{d+\log(1/\beta)}{\alpha^2}\right)$. Here $f(D) := \frac{1}{n}\sum_{x \in D} f(x)$ and $f(P) := \mathbf{E}_{X \leftarrow P}[f(X)]$.

It is immediate from Definition IV.2 that $VC(\mathcal{F}) \leq \lfloor \log_2 |\mathcal{F}| \rfloor$. Thus Theorem IV.4 subsumes the union bound used in the proof of Theorem I.1.

The relevant application of uniform convergence for our algorithm is the following lemma (roughly the equivalent of Lemma III.1), which says that good hypotheses have high scores, and bad hypotheses have low scores.

Lemma IV.5. *Let $\mathcal{H}$ be a collection of probability distributions on $\mathcal{X}$ with VC dimension d.*

Let $S : \mathcal{H} \times \mathcal{X}^n \to \mathbb{R}$ be a score function similar to (1), namely $S(H, D) = \inf_{H' \in \mathcal{H}} \max\{|\{x \in D : H(x) > H'(x)\}| - n \cdot (\Pr_{X \leftarrow H'}[H(X) > H'(X)] + 3\alpha), n \cdot \mathbb{I}[d_{\mathrm{TV}}(H, H') \leq 6\alpha]\}$, where $\mathbb{I}$ denotes the indicator function.

Let P be a distribution on $\mathcal{X}$. Let $\alpha, \beta > 0$ and $n \geq O(\frac{1}{\alpha^2}(d + \log(1/\beta)))$. Suppose there exists $H^* \in \mathcal{H}$ with $d_{\mathrm{TV}}(P, H^*) \leq \alpha$. Then, with probability at least $1 - \beta$ over $D \leftarrow P^n$, we have

- *$S(H^*, D) > \alpha n$ and*
- *$S(H, D) = 0$ for all $H \in \mathcal{H}$ with $d_{\mathrm{TV}}(H, P) > 7\alpha$.*

Proof. For $H, H' \in \mathcal{H}$, define $f_{H,H'} : \mathcal{X} \to \{0,1\}$ by $f_{H,H'}(x) = 1 \iff H(x) > H'(x)$. Note that $|\{x \in D : H(x) > H'(x)\}| = \sum_{x \in D} f_{H,H'}(x)$ and d is the VC dimension of the function class $\{f_{H,H'} : H, H' \in \mathcal{H}\}$. By Theorem IV.4, if $n = \Omega\left(\frac{d+\log(1/\beta)}{\alpha^2}\right)$, then $\Pr_{D \leftarrow P^n}[\forall H, H' \in \mathcal{H} \ \ ||\{x \in D : H(x) > H'(x)\}| - n \cdot \Pr_{X \leftarrow P}[H(X) > H'(X)]| \leq \alpha n] \geq 1 - \beta$ We condition on this event happening.

In order to prove the first conclusion – namely, $S(H^*, D) > \alpha n$ – it remains to show that, for all $H' \in \mathcal{H}$, we have either $d_{\mathrm{TV}}(H^*, H') \leq 6\alpha$ or $|\{x \in D : H(x) > H'(x)\}| - n \cdot (\Pr_{X \leftarrow H'}[H^*(X) > H'(X)] + 3\alpha) > \alpha n$. If $d_{\mathrm{TV}}(H^*, H') \leq 6\alpha$, we are done, so assume $d_{\mathrm{TV}}(H^*, H') > 6\alpha$. By the uniform convergence event we have conditioned on,

$$\begin{aligned}
&|\{x \in D : H(x) > H'(x)\}| \\
&\geq n \cdot (\Pr_{X \leftarrow P}[H(X) > H'(X)] - \alpha) \\
&\geq n \cdot (\Pr_{X \leftarrow H^*}[H(X) > H'(X)] - d_{\mathrm{TV}}(P, H^*) - \alpha) \\
&\geq n \cdot (d_{\mathrm{TV}}(H^*, H') + \Pr_{X \leftarrow H'}[H(X) > H'(X)] - 2\alpha) \\
&> n \cdot (6\alpha + \Pr_{X \leftarrow H'}[H(X) > H'(X)] - 2\alpha),
\end{aligned}$$

from which the desired conclusion follows.

In order to prove the second conclusion – namely, $S(H, D) = 0$ for all $H \in \mathcal{H}$ with $d_{\mathrm{TV}}(H, P) > 7\alpha$ – it suffices to show that one $H' \in \mathcal{H}$ yields a score of zero for any $H \in \mathcal{H}$ with $d_{\mathrm{TV}}(H, P) > 7\alpha$. In particular, we show that $H' = H^*$ yields a score of zero for any such H. That is, if $d_{\mathrm{TV}}(H, P) > 7\alpha$, then $d_{\mathrm{TV}}(H, H^*) > 6\alpha$ and $|\{x \in D : H(x) > H^*(x)\}| - n \cdot (\Pr_{X \leftarrow H^*}[H(X) > H^*(X)] + 3\alpha) \leq 0$. By the triangle inequality $d_{\mathrm{TV}}(H, H^*) \geq d_{\mathrm{TV}}(H, P) - d_{\mathrm{TV}}(P, H^*) > 7\alpha - \alpha = 6\alpha$, as required. By the uniform convergence event we have conditioned on,

$$\begin{aligned}
&|\{x \in D : H(x) > H^*(x)\}| \\
&\leq n \cdot (\Pr_{X \leftarrow P}[H(X) > H^*(X)] + \alpha) \\
&\leq n \cdot (\Pr_{X \leftarrow H^*}[H(X) > H^*(X)] + d_{\mathrm{TV}}(P, H^*) + \alpha) \\
&\leq n \cdot (\Pr_{X \leftarrow H^*}[H(X) > H^*(X)] + 2\alpha),
\end{aligned}$$

which completes the proof. □

B. GAP-MAX Algorithm

In place of the exponential mechanism for privately selecting a hypothesis we use the following algorithm

[5] Here, for simplicity, we assume that each distribution H is given by a density function $H(\cdot)$. More generally, we define the VC dimension of $\mathcal{H}$ to be the smallest d such that there exists a function family $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ of VC dimension d with the property that, for all $H, H' \in \mathcal{H}$ we have $d_{\mathrm{TV}}(H, H') = \sup_{f \in \mathcal{F}} \mathbf{E}_{X \leftarrow H}[f(X)] - \mathbf{E}_{X \leftarrow H'}[f(X)]$, where the supremum is over f measurable with respect to both H and H'. We ignore this technicality throughout.

that works under a "gap" assumption. That is, we assume that there is a $5\alpha n$ gap between the highest score and the $(k+1)$-th highest score. Rather than paying in sample complexity for the total number of hypotheses we pay for the number of high-scoring hypotheses k.

This algorithm is based on the GAP-MAX algorithm of Bun, Dwork, Rothblum, and Steinke [23]. However, we combine their GAP-MAX algorithm with the exponential mechanism to improve the dependence on the parameter k.

Theorem IV.6. *Let $\mathcal{H}$ and $\mathcal{X}$ be arbitrary sets. Let $S : \mathcal{H} \times \mathcal{X}^n \to \mathbb{R}$ have sensitivity at most 1 in its second argument – that is, for all $H \in \mathcal{H}$ and all $D, D' \in \mathcal{X}^n$ differing in a single example, $|S(H,D) - S(H,D')| \leq 1$.*

For $D \in \mathcal{X}^n$ and $\alpha > 0$, define $K(D, 5\alpha) := |\{H \in \mathcal{H} : S(H,D) \geq \sup_{H' \in \mathcal{H}} S(H', D) - 5\alpha n\}|$.

Given parameters $\varepsilon, \delta, \beta > 0$ and $n, k \geq 1$, there exists a (ε, δ)-differentially private randomized algorithm $M : \mathcal{X}^n \to \mathcal{H}$ such that, for all $D \in \mathcal{X}^n$ and all $\alpha > 0$, $K(D, 5\alpha) \leq k \implies \Pr[S(M(D), D) \geq \sup_{H' \in \mathcal{H}} S(H', D) - \alpha n] \geq 1 - \beta$ provided $n = \Omega\left(\frac{\min\{\log|\mathcal{H}|, \log(1/\delta)\} + \log(k/\beta)}{\alpha\varepsilon}\right)$.

Furthermore, given $\varepsilon, \beta > 0$ and $n, k \geq 1$, there exists a $\frac{1}{2}\varepsilon^2$-concentrated differentially private [94] algorithm $M : \mathcal{X}^n \to \mathcal{H}$ such that, for all $D \in \mathcal{X}^n$ and all $\alpha > 0$, $K(D, 5\alpha) \leq k \implies \Pr[S(M(D), D) \geq \sup_{H' \in \mathcal{H}} S(H', D) - \alpha n] \geq 1 - \beta$ provided $n = \Omega\left(\frac{\sqrt{\log|\mathcal{H}| + \log(k/\beta)}}{\alpha\varepsilon}\right)$.

Proof. We begin by describing the algorithm.

1) Let $m = \left\lceil \frac{k^2}{\beta} \right\rceil$ and let $G : \mathcal{H} \to [m]$ be a uniformly random function.[6]
2) Randomly select $B \in [m]$ with $\Pr[B = b] \propto \exp\left(\frac{\varepsilon}{4} \sup\{S(H,D) : H \in \mathcal{H}, G(H) = b\}\right)$.
3) Define $\mathcal{H}_B = \{H \in \mathcal{H} : G(H) = B\}$. Let $H_B^1 = \operatorname{argmax}_{H \in \mathcal{H}_B} S(H,D)$ and $H_B^2 = \operatorname{argmax}_{H \in \mathcal{H}_B \setminus \{H_B^1\}} S(H,D)$, breaking ties arbitrarily. (That is, $\mathcal{H}_B$ is the B-th "bin" and H_B^1 and H_B^2 are the items in this bin with the largest and second-largest scores respectively.) Define $S'_B : \mathcal{H}_B \times \mathcal{X}^n \to \mathbb{R}$ by
$$S'_B(H,D) = \frac{1}{2}\max\{0, S(H,D) - S(H_B^2, D)\}.$$
(Note that S'_B has sensitivity 1 and $S'_B(H,D) = 0$ whenever $H \neq H_B^1$.)
4) Let $\mathcal{D}$ be a distribution on $\mathbb{R}$ such that adding a sample from $\mathcal{D}$ to a sensitivity-1 function provides $(\varepsilon/4, \delta/2)$-differential privacy (or, respectively, $\frac{1}{6}\varepsilon^2$-concentrated differential privacy). For example, $\mathcal{D}$ could be a Laplace distribution with scale $4/\varepsilon$ truncated to the interval $[-t, t]$ for $t = 4(1 + \log(1/\delta))/\varepsilon$ (or unbounded if $\delta = 0$). To attain concentrated differential privacy, we can set $\mathcal{D} = N\left(0, \frac{3}{\varepsilon^2}\right)$, a centered Gaussian with variance $3/\varepsilon^2$.
5) Draw a sample Z_H i.i.d. from $\mathcal{D}$ corresponding to every $H \in \mathcal{H}_B$.
6) Return $H^* = \operatorname{argmax}_{H \in \mathcal{H}_B} S'_B(H,D) + Z_H$.

The selection of B is an instantiation of the exponential mechanism [20] and is $(\varepsilon/2, 0)$-differentially private. The selection of H^* in the final step is a GAP-MAX algorithm [23] and is $(\varepsilon/2, \delta)$-differentially private. By composition, the entire algorithm is (ε, δ)-differentially private (or, respectively, $\frac{1}{2}\varepsilon^2$-concentrated differentially private).

For the utility analysis, in order for the algorithm to output a good H^*, it suffices for the following three events to occur.

- $S(H_B^1, D) \geq \sup_{H' \in \mathcal{H}} S(H', D) - \alpha n$.
  That is, restricting the search to $\mathcal{H}_B$, rather than all of $\mathcal{H}$, only reduces the score of the optimal choice by αn. The exponential mechanism ensures that this happens with probability at least $1 - \beta/4$, as long as $n \geq \frac{4\log(2k/\beta)}{\varepsilon\alpha}$.
- $S(H_B^2, D) < \sup_{H' \in \mathcal{H}} S(H', D) - 5\alpha n$.
  That is, the second-highest score within $\mathcal{H}_B$ is at least $5\alpha n$ less than the highest score overall. We have assumed that there are at most k elements $H \in \mathcal{H}$ such that $S(H,D) \geq \sup_{H' \in \mathcal{H}} S(H', D) - 5\alpha n$. Call these "large elements." Since $G : \mathcal{H} \to [m]$ is random and $m \geq k^2/\beta$, the probability that an arbitrary but fixed pair of large elements collide – that is, are in the same H_B is $1/m \leq \beta/k^2$. If we union bound over the $\binom{k}{2} < k^2/2$ pairs, we see that the probability of any collisions is at most $\beta/2$. Thus, the probability that more than one large element satisfies $G(H) = B$ is at most $\beta/2$. This suffices for the event to occur.
- $\sup_{H \in \mathcal{H}_B} |Z_H| \leq \alpha n$.
  If the noise distribution $\mathcal{D}$ is supported on $[-\alpha n, \alpha n]$, then this condition holds with probability 1. For the truncated Laplace distribution, this is possible whenever $n \geq 1 + 4\log(1/\delta)/\alpha\varepsilon$. Alternatively, we can use unbounded Laplace noise and a union bound to show that this event occurs with probability at least $1 - \beta/4$ whenever $n \geq 4\log(4|\mathcal{H}_B|/\beta)/\varepsilon\alpha$. For Gaussian noise, $n \geq \frac{3}{\varepsilon\alpha}\sqrt{\log(4|\mathcal{H}_B|/\beta)}$ suffices.

Assuming the first and second events occur, we have $S'_B(H_B^1, D) = \frac{S(H_B^1, D) - S(H_B^2, D)}{2} > 2\alpha n$. Given this, the third event implies $H^* = H_B^1$. Finally, the first event then implies $S(H^*, D) \geq \sup_{H' \in \mathcal{H}} S(H', D) - \alpha n$, as

[6] It suffices for G to be a drawn from a universal hash function family.

required. A union bound over the three events completes the proof. □

Now we can combine the VC-based uniform convergence bound with the GAP-MAX algorithm to prove our result.

Proof of Theorem IV.1. By Lemma IV.5, with high probability over the draw of the dataset D, our score function satisfies $\sup_{H\in\mathcal{H}} S(H,D) \geq S(H^*,D) > \alpha n$ and $S(H,D) = 0$ whenever $d_{\mathrm{TV}}(H,P) > 7\alpha$. This requires $n = \Omega(d/\alpha^2)$.

Note that the score function S has sensitivity-1, since it is the supremum of counts. Conditioned on the uniform convergence event, the maximum score is at least αn and there are at most k elements of $\mathcal{H}$ with score greater than 0. Thus we can apply the GAP-MAX algorithm of Theorem IV.6. If $n = \Omega((\min\{\log|\mathcal{H}|, \log(1/\delta)\} + \log(k))/\alpha\varepsilon)$, then with high probability, the algorithm outputs $\hat{H} \in \mathcal{H}$ with score at least $\frac{4}{5}\alpha n$, as required. □

V. Packings, Lower Bounds, and Relations to Covers

In this section, we show that the sample complexity of our algorithms for private hypothesis selection with pure differential privacy cannot be improved, at least for constant values of the proximity parameter α. We first apply a packing argument [98], [99] to show a lower bound which is logarithmic in the packing number of the class of distributions (Lemma V.1). We then state a folklore relationship between the sizes of maximal packings and minimal covers (Lemma V.2), which shows that instantiating our private hypothesis selection algorithm with a minimal cover gives essentially optimal sample complexity (Theorem V.3).

Lemma V.1. *Suppose there exists an α-packing $\mathcal{P}_\alpha$ of a set of distributions $\mathcal{H}$. Then any ε-differentially private algorithm which takes as input samples $X_1,\ldots,X_n \sim P$ for some $P \in \mathcal{H}$ and produces a distribution $\hat{H}$ such that $d_{\mathrm{TV}}(P,\hat{H}) \leq \alpha/2$ with probability $\geq 9/10$ requires*

$$n = \Omega\left(\frac{\log|\mathcal{P}_\alpha|}{\varepsilon}\right).$$

One might conjecture a stronger version of this lemma exists, and that one could prove the lower bound $n = \Omega\left(\frac{\log|\mathcal{P}_\alpha|}{\alpha\varepsilon}\right)$. However, such a statement cannot be true in general. For example, consider an α-packing of $N(\mu,1)$ where $\mu \in [-R,R]$, which would have size $\Omega(R/\alpha)$. If such a lemma were true, it would imply a lower bound of $\tilde{\Omega}\left(\frac{\log R}{\alpha\varepsilon}\right)$, which contradicts known upper bounds. One way to prove a lower bound achieving such a dependence on α would be to have a single "central" distribution which is close to all distributions in the packing (see an argument of this sort in Theorem 5.13 of [100]). However, we do not explore this here, as our goal is to match our upper bound which is stated in terms of a generic cover.

Proof. Let M be a ε-differentially private algorithm with the stated accuracy requirement, and denote by $M(P^n)$ the distribution on hypotheses obtained by running M on n i.i.d. samples from a distribution $P \in \mathcal{H}$. For each $P \in \mathcal{P}_\alpha$, let B_P denote the set of distributions which are at total variation distance at most $\alpha/2$ from P. Then the accuracy requirement implies that $\Pr_{\hat{H}\leftarrow M(P^n)}\left[\hat{H} \in B_P\right] \geq 9/10$ for all $P \in \mathcal{H}$. Let $P_0 \in \mathcal{P}_\alpha$ be an arbitrary packing element. Note that, trivially, samples from P^n and P_0^n have Hamming distance at most n for any P. Recall the group privacy property of differential privacy, which states that if M is ε-DP, then $\Pr[M(X) \in S] \leq \exp(\varepsilon d(X,X')) \cdot \Pr[M(X') \in S]$ for any set $S \subseteq \mathrm{Range}(M)$, where $d(X,X')$ is the Hamming distance between the two datasets. Applying this property with P^n and P_0^n, we have

$$\Pr_{\hat{H}\leftarrow M(P_0^n)}\left[\hat{H} \in B_P\right] \geq e^{-\varepsilon n} \cdot 9/10$$

for every $P \in \mathcal{P}_\alpha$. The fact that $\mathcal{P}_\alpha$ is an α-packing implies that the sets B_P are all disjoint, and hence

$$1 \geq \sum_{P\in\mathcal{P}_\alpha} \Pr_{\hat{H}\leftarrow M(P_0^n)}\left[\hat{H} \in B_P\right] \geq |\mathcal{P}_\alpha| \cdot e^{-\varepsilon n} \cdot 9/10.$$

Rearranging gives us the stated lower bound on n. □

The following lemma is a well-known folklore relationship between packing and covering numbers. We include a proof for completeness.

Lemma V.2. *For a set of distributions $\mathcal{H}$, let p_α and c_α be the size of the largest α-packing and smallest α-cover of $\mathcal{H}$, respectively. Then*

$$p_{2\alpha} \leq c_\alpha \leq p_\alpha.$$

Proof. We first prove the inequality on the left. Let $\mathcal{C}_\alpha$ be a cover of $\mathcal{H}$ of minimal size c_α. If $c_\alpha = \infty$, we are done. Otherwise, let $S \subseteq \mathcal{H}$ be any set of distributions of size at least $c_\alpha + 1$. By the pigeonhole principle, there exists $P \in \mathcal{C}_\alpha$ and two distinct distributions $Q, Q' \in S$ such that $d_{\mathrm{TV}}(P,Q) \leq \alpha$ and $d_{\mathrm{TV}}(P,Q') \leq \alpha$. Hence $d_{\mathrm{TV}}(Q,Q') \leq 2\alpha$ by the triangle inequality, so S cannot be (2α)-packing of $\mathcal{H}$. This suffices to show that $p_{2\alpha} \leq c_\alpha$.

Next, we prove the inequality on the right. Let $\mathcal{P}_\alpha$ be a maximal α-packing with size $|\mathcal{P}_\alpha| = p_\alpha$. If $p_\alpha = \infty$, we are done. Otherwise, we claim that $\mathcal{P}_\alpha$ is also an α-cover of $\mathcal{H}$, and hence $c_\alpha \leq |\mathcal{P}_\alpha| = p_\alpha$. To see this, suppose for the sake of contradiction that there were a distribution $P \in \mathcal{H}$ with $d_{\mathrm{TV}}(P,\mathcal{P}_\alpha) > \alpha$. Then we could add P to $\mathcal{P}_\alpha$ to produce a strictly larger packing, contradicting the maximality of $\mathcal{P}_\alpha$. □

Theorem V.3. *Let $\mathcal{H}$ be a set of distributions, and let n^*_α denote the minimum number of samples such that there exists an ε-differentially private algorithm which takes as input samples $X_1, \ldots, X_{n^*_\alpha} \sim P$ for an arbitrary $P \in \mathcal{H}$ and outputs a distribution $\hat{H}$ such that $d_{\mathrm{TV}}(P, \hat{H}) \leq \alpha/2$ with probability $\geq 9/10$. Then there exists a cover of $\mathcal{H}$ such that the instantiation of the algorithm underlying Theorem I.1 with this cover takes as input $n = \Omega(n^*_\alpha \cdot (\varepsilon/\alpha^2 + 1/\alpha))$ samples from an arbitrary $P \in \mathcal{H}$ and outputs a $\hat{H}$ such that $d_{\mathrm{TV}}(P, \hat{H}) \leq (6 + 2\zeta)\alpha$ with probability $\geq 9/10$ for any constant $\zeta > 0$.*

Proof. Let p_α denote the size of the largest α-packing of $\mathcal{H}$. By Lemma V.1, we have $n^*_\alpha = \Omega(\log p_\alpha/\varepsilon)$. On the other hand, by Lemma V.2, we know that there exists an α-cover $\mathcal{C}_\alpha$ of $\mathcal{H}$ with $|\mathcal{C}_\alpha| \leq p_\alpha$. Hence $\log |\mathcal{C}_\alpha| \leq O(\varepsilon \cdot n^*_\alpha)$ and the asserted sample complexity guarantee follows from Corollary I.2. □

VI. Applications of Hypothesis Selection

In this section, we give a number of applications of Theorem I.1, primarily to obtain sample complexity bounds for learning a number of distribution classes of interest. Recall Corollary I.2, which is an immediate corollary of Theorem I.1. This indicates that we can privately semi-agnostically learn a class of distributions with a number of samples proportional to the logarithm of its covering number.

Corollary I.2. *Suppose there exists an α-cover $\mathcal{C}_\alpha$ of a set of distributions $\mathcal{H}$, and that we are given a set of samples $X_1, \ldots, X_n \sim P$, where $d_{\mathrm{TV}}(P, \mathcal{H}) \leq \alpha$. For any constant $\zeta > 0$, there exists an ε-differentially private algorithm (with respect to the input $\{X_1, \ldots, X_n\}$) which outputs a distribution $H^* \in \mathcal{C}_\alpha$ such that $d_{\mathrm{TV}}(P, H^*) \leq (6 + 2\zeta)\alpha$ with probability $\geq 9/10$, as long as*

$$n = \Omega\left(\frac{\log |\mathcal{C}_\alpha|}{\alpha^2} + \frac{\log |\mathcal{C}_\alpha|}{\alpha\varepsilon}\right).$$

Note that the factor of $(6+2\zeta)\alpha$ in the corollary statement (versus $(3+\zeta)\alpha$ in the statement of Theorem I.1) is due to the fact the algorithm is semi-agnostic, and the closest element in the cover is 2α-close to P, rather than just α-close.

We instantiate this result to give the sample complexity results for semi-agnostically learning product distributions (Section VI-A), Gaussian distributions (Section VI-B), sums of some independent random variable classes (Section VI-C), piecewise polynomials (Section VI-D), and mixtures (Section VI-E). Furthermore, we mention an application to private PAC learning (Section VI-F), when the distribution of unlabeled examples is known to come from some hypothesis class.

A. Product Distributions

As a first application, we first give an ε-differentially private algorithm for learning product distributions over discrete alphabets.

Definition VI.1. *A* (k, d)-product distribution *is a distribution over $[k]^d$, such that its marginal distributions are independent (i.e., the distribution is the product of its marginals).*

We start by constructing a cover for product distributions.

Lemma VI.2. *There exists an α-cover of the set of (k, d)-product distributions of size*

$$O\left(\frac{kd}{\alpha}\right)^{d(k-1)}.$$

Proof. Consider some fixed product distribution P, with marginal distributions $(P_1, \ldots, P_d)$. We will construct a cover that contains a distribution Q (with marginals $(Q_1, \ldots, Q_d)$) that is α-close in total variation distance.

First, by triangle inequality, we have that $d_{\mathrm{TV}}(P, Q) \leq \sum_{i=1}^d d_{\mathrm{TV}}(P_i, Q_i)$, so it suffices to approximate each marginal distribution to accuracy α/d. Stated another way, we must generate an (α/d)-cover of distributions over $[k]$, and we can then take its d-wise Cartesian product. Raising the size of this underlying cover to the power d gives us the size of the overall cover.

To (α/d)-cover a distribution over $[k]$, we will additively grid the probability of each symbol at granularity $\Theta\left(\frac{\alpha}{kd}\right)$, choosing the probability of the last symbol k such that the sum is normalized. This will incur $\Theta\left(\frac{\alpha}{kd}\right)$ error per symbol (besides for symbol k), and summing over the $k-1$ symbols accumulates error $\Theta\left(\frac{\alpha}{d}\right)$. It can also be argued that the error on symbol k is $O\left(\frac{\alpha}{d}\right)$ – with an appropriate choice of granularity, this gives us an (α/d)-cover for distributions over $[k]$. The size of this cover is $O\left(\frac{kd}{\alpha}\right)^{k-1}$, which allows us to conclude the lemma statement. □

With this cover in hand, applying Corollary I.2 allows us to conclude the following sample complexity upper bound.

Corollary VI.3. *Suppose we are given a set of samples $X_1, \ldots, X_n \sim P$, where P is α-close to a (k, d)-product distribution. Then for any constant $\zeta > 0$, there exists an ε-differentially private algorithm which outputs a (k, d)-product distribution H^* such that $d_{\mathrm{TV}}(P, H^*) \leq (6 + 2\zeta)\alpha$ with probability $\geq 9/10$, so long as*

$$n = \Omega\left(kd \log\left(\frac{kd}{\alpha}\right)\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)\right).$$

This gives the first $\tilde{O}(d)$ sample algorithm for learning a binary product distribution in total variation distance under pure differential privacy, improving upon the work of Kamath, Li, Singhal, and Ullman [17] by strengthening the privacy guarantee at a minimal cost in the sample complexity. The natural way to adapt their result from concentrated to pure differential privacy would require $\Omega(d^{3/2})$ samples.

Remark VI.4. *Properly learning a product distribution over $\{0,1\}^d$ to total variation distance $\leq \frac{1}{2}$ implies learning its mean $\mu \in [0,1]^d$ up to ℓ_1 error $\leq 2\sqrt{d}$; see Lemma VI.5 below.*

Thus Corollary VI.3 implies a ε-differentially private algorithm which takes $n = \tilde{O}(d/\varepsilon)$ samples from a product distribution P on $\{0,1\}^d$ and, with high probability, outputs an estimate $\hat{\mu}$ of its mean μ with $\|\hat{\mu} - \mu\|_1 \leq 2\sqrt{d}$.

In contrast, for non-product distributions over the hypercube, estimating the mean to the same accuracy under ε-differential privacy requires $n = \Omega(d^{3/2}/\varepsilon)$ samples [98], [101]. Thus we have a polynomial separation between estimating product and non-product distributions under pure differential privacy.

Lemma VI.5. *If P and Q are product distributions on $\mathbb{R}^d$ with $d_{\mathrm{TV}}(P,Q) \leq \frac{1}{2}$ and per-coordinate variance at most σ^2, then*

$$\left\|\mathbf{E}_{X \leftarrow P}[X] - \mathbf{E}_{X \leftarrow Q}[X]\right\|_1 \leq 4\sqrt{d\sigma^2}.$$

Proof. Let $\mu = \mathbf{E}_{X \leftarrow P}[X] \in \mathbb{R}^d$ and $\mu' = \mathbf{E}_{X \leftarrow Q}[X] \in \mathbb{R}^d$. Let $\tau = \|\mu - \mu'\|_1$. Let $\nu = \mathrm{sign}(\mu - \mu') \in \{-1,+1\}^d$ so that $\langle \nu, \mu - \mu' \rangle = \tau$. We have $\frac{1}{2} \geq d_{\mathrm{TV}}(P,Q) \geq \Pr_{X \leftarrow P}[\langle \nu, X \rangle \geq t] - \Pr_{X \leftarrow Q}[\langle \nu, X \rangle \geq t] = \Pr_{X \leftarrow P}[\langle \nu, X - \mu \rangle \geq t - \langle \nu, \mu \rangle] - \Pr_{X \leftarrow Q}[\langle \nu, X - \mu' \rangle \geq t - \langle \nu, \mu \rangle + \langle \nu, \mu - \mu' \rangle]$. We set $t = \langle \nu, \mu \rangle - \frac{\tau}{2}$, and this is equal to $\Pr_{X \leftarrow P}\left[\langle \nu, X - \mu \rangle \geq -\frac{\tau}{2}\right] - \Pr_{X \leftarrow Q}\left[\langle \nu, X - \mu' \rangle \geq +\frac{\tau}{2}\right]$. Chebyshev's inequality implies that this is $\geq 1 - \frac{\mathbf{E}_{X \leftarrow P}\left[\langle \nu, X - \mu \rangle^2\right]}{(\tau/2)^2} - \frac{\mathbf{E}_{X \leftarrow Q}\left[\langle \nu, X - \mu' \rangle^2\right]}{(\tau/2)^2} = 1 - \frac{4}{\tau^2}\sum_{i=1}^{d} \mathbf{E}_{X \leftarrow P}\left[(X_i - \mu_i)^2\right] + \mathbf{E}_{X \leftarrow Q}\left[(X_i - \mu_i')^2\right] \geq 1 - \frac{8d\sigma^2}{\tau^2}$. Rearranging yields $\tau \leq 4\sqrt{d\sigma^2}$, as required. □

B. Gaussian Distributions

We next give private algorithms for learning Gaussian distributions.

Definition VI.6. *A* Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ *in $\mathbb{R}^d$ is a distribution with PDF*

$$p(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^d |\Sigma|}}.$$

We describe covers for Gaussian distributions with known and unknown covariance.

Lemma VI.7. *There exists an α-cover of the set of Gaussian distributions $\mathcal{N}(\mu, I)$ in d dimensions with $\|\mu\|_2 \leq R$ of size*

$$O\left(\frac{dR}{\alpha}\right)^d.$$

Proof. It is well-known that estimating a Gaussian distribution with unknown mean in total variation distance corresponds to estimating μ in ℓ_2-distance (see, e.g., [10]). By the triangle inequality, in order to α-cover the space, it suffices to (α/d)-cover each standard basis direction. Since we know the mean in each direction is bounded by R, a simple additive grid in each direction with granularity $\Theta\left(\frac{\alpha}{d}\right)$ will suffice, resulting in a cover for each direction of size $O\left(\frac{dR}{\alpha}\right)$. Taking the Cartesian product over d dimensions gives the desired result. □

Lemma VI.8. *There exists an α-cover of the set of Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ in d-dimensions with $\|\mu\|_2 \leq R$ and $I \preceq \Sigma \preceq \kappa I$ of size*

$$O\left(\frac{dR}{\alpha}\right)^d \cdot O\left(\frac{d\kappa}{\alpha}\right)^{d(d+1)/2}.$$

Proof. The former term is obtained similarly to the expression in Lemma VI.7. Since $I \preceq \Sigma$, we can still bound the total variation contribution by the ℓ_2-distance between the mean vectors. We thus turn our attention to the latter term. To construct our cover, we must argue about the total variation distance between $\mathcal{N}(0, \Sigma)$ and $\mathcal{N}(0, \hat{\Sigma})$. If $|\Sigma(i,j) - \hat{\Sigma}(i,j)| \leq \gamma$, and $I \preceq \Sigma$, Proposition 32 of [102] implies:

$$d_{\mathrm{TV}}(\mathcal{N}(0, \Sigma), \mathcal{N}(0, \hat{\Sigma})) \leq O(d\gamma).$$

We will thus perform a gridding, in order to approximate each entry of Σ to an additive $O(\gamma) = O(\alpha/d)$. However, in order to ensure that the resulting matrix is PSD, we grid over entries of $\hat{\Sigma}$'s Cholesky decomposition, rather than grid for $\hat{\Sigma}$ itself. Since the largest element of Σ is bounded by κ, the larest element of its Cholesky decomposition must be bounded by $\sqrt{\kappa}$. An additive grid over the range $[0, \sqrt{\kappa}]$ with granularity $O(\gamma/\sqrt{\kappa})$ suffices to get $\hat{\Sigma}$ which bounds the entrywise distance as $O(\gamma)$. This requires $O(d\kappa/\alpha)$ candidates per entry, and we take the Cartesian product over all $d(d+1)/2$ entries of the Cholesky decomposition, giving the desired result. □

In addition, we can obtain bounds of the VC dimension of the Scheffé sets of Gaussian distributions.

Lemma VI.9. *The set of Gaussian distributions with fixed variance – i.e., all $\mathcal{N}(\mu, I)$ with $\mu \in \mathbb{R}^d$ – has VC dimension $d+1$. Furthermore, the set of Gaussians with* *unknown variance – i.e., all $\mathcal{N}(\mu,\Sigma)$ with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$ positive definite – has VC dimension $O(d^2)$.*

Proof. For Gaussians with fixed variance, the Scheffé sets correspond to linear threshold functions, which have VC dimension $d+1$. For Gaussians with unknown variance, the Scheffé sets correspond to quadratic threshold functions, which have VC dimension $\binom{d+2}{2} = O(d^2)$ [103]. □

Combining the covers of Lemmas VI.7 and VI.8 and the VC bound of Lemma VI.9 with Theorem IV.1 implies the following corollaries for Gaussian estimation.

Corollary VI.10. *Suppose we are given a set of samples $X_1,\ldots,X_n \sim P$, where P is α-close to a Gaussian distribution $\mathcal{N}(\mu,I)$ in d-dimensions with $\|\mu\| \le R$. Then for any constant $\zeta > 0$, there exists an ε-differentially private algorithm which outputs a Gaussian distribution H^* such that $d_{\mathrm{TV}}(P,H^*) \le (6+2\zeta)\alpha$ with probability $\ge 9/10$, so long as*

$$n = \Omega\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon}\log\left(\frac{dR}{\alpha}\right)\right).$$

Corollary VI.11. *Suppose we are given a set of samples $X_1,\ldots,X_n \sim P$, where P is α-close to a Gaussian distribution $\mathcal{N}(\mu,\Sigma)$ in d-dimensions with $\|\mu\| \le R$ and $I \preceq \Sigma \preceq \kappa I$. Then for any constant $\zeta > 0$, there exists an ε-differentially private algorithm which outputs a Gaussian distribution H^* such that $d_{\mathrm{TV}}(P,H^*) \le (6+2\zeta)\alpha$ with probability $\ge 9/10$, so long as*

$$n = \Omega\left(\frac{d^2}{\alpha^2} + \frac{1}{\alpha\varepsilon}\left(d\log\left(\frac{dR}{\alpha}\right) + d^2\log\left(\frac{d\kappa}{\alpha}\right)\right)\right).$$

Similar to the product distribution case, these are the first $\tilde{O}(d)$ and $\tilde{O}(d^2)$ sample algorithms for learning Gaussians total variation distance under pure differential privacy, improving upon the concentrated differential privacy results of Kamath, Li, Singhal, and Ullman [17].

1) Gaussians with Unbounded Mean: Extending Corollary VI.10, we consider multivariate Gaussian hypotheses with known covariance and unknown mean, *without* assuming bound on the mean (the parameter R in the discussion above). To handle the unbounded mean we must relax to approximate differential privacy.

In place of Lemma VI.7, we construct a locally small cover:

Lemma VI.12. *For any $d \in \mathbb{N}$ and $\alpha \in (0,1/30]$, there exists an α-cover $\mathcal{C}_\alpha$ of the set of Gaussian distributions $\mathcal{N}(\mu,I)$ in d dimensions satisfying*

$$\forall \mu \in \mathbb{R}^d \quad |\{H \in \mathcal{C}_\alpha : d_{\mathrm{TV}}(H,\mathcal{N}(\mu,I)) \le 7\alpha\}| \le 2^{15d}.$$

Proof. For $\mu,\mu' \in \mathbb{R}^d$, we have

$$\begin{aligned} &d_{\mathrm{TV}}(\mathcal{N}(\mu,I),\mathcal{N}(\mu',I)) \\ &= 2\Pr\left[\mathcal{N}(0,1) \in \left[0,\frac{1}{2}\|\mu-\mu'\|_2\right]\right] \\ &= \sqrt{\frac{2}{\pi}}\int_0^{\frac{1}{2}\|\mu-\mu'\|_2} e^{-x^2/2}\mathrm{d}x \\ &\le \frac{\|\mu-\mu'\|_2}{\sqrt{2\pi}}. \end{aligned}$$

Furthermore, for any $c > 0$, $d_{\mathrm{TV}}(\mathcal{N}(\mu,I),\mathcal{N}(\mu',I)) \ge \begin{cases} \frac{\|\mu-\mu'\|_2}{\sqrt{2\pi}}\cdot e^{-c^2/2} & \text{if } \frac{1}{2}\|\mu-\mu'\|_2 \le c \\ \frac{c\cdot e^{-c^2/2}}{\sqrt{2\pi}} & \text{if } \frac{1}{2}\|\mu-\mu'\|_2 \ge c \end{cases}$.

$$\text{Let } \mathcal{C}_\alpha = \left\{\mathcal{N}\left(m\cdot\frac{\alpha\sqrt{8\pi}}{\sqrt{d}},I\right) : m \in \mathbb{Z}^d\right\}.$$

Fix $\mu \in \mathbb{R}^d$. Let $\mu^* = \mu\frac{\sqrt{d}}{\alpha\sqrt{8\pi}} \in \mathbb{R}^d$ and let $m \in \mathbb{Z}^d$ be μ^* rounded to the nearest integer coordinate-wise, so that $\|m-\mu^*\|_\infty \le \frac{1}{2}$. Then

$$\begin{aligned} &d_{\mathrm{TV}}\left(\mathcal{N}(\mu,I),\mathcal{N}\left(m\cdot\frac{\alpha\sqrt{8\pi}}{\sqrt{d}},I\right)\right) \\ &= d_{\mathrm{TV}}\left(\mathcal{N}\left(\mu^*\cdot\frac{\alpha\sqrt{8\pi}}{\sqrt{d}},I\right),\mathcal{N}\left(m\cdot\frac{\alpha\sqrt{8\pi}}{\sqrt{d}},I\right)\right) \\ &\le \frac{1}{\sqrt{2\pi}}\frac{\alpha\sqrt{8\pi}}{\sqrt{d}}\|\mu^*-m\|_2 \\ &\le \alpha, \end{aligned}$$

since $\|\mu^*-m\|_2 \le \sqrt{d}\|\mu^*-m\|_\infty \le \frac{\sqrt{d}}{2}$. This proves that $\mathcal{C}_\alpha$ is a α-cover of $\{\mathcal{N}(\mu,I) : \mu \in \mathbb{R}^d\}$.

It remains to show that the cover is "locally small". Let $m' \in \mathbb{Z}^d$. Then

$$\begin{aligned} &d_{\mathrm{TV}}\left(\mathcal{N}(\mu,I),\mathcal{N}\left(m'\cdot\frac{\alpha\sqrt{8\pi}}{\sqrt{d}},I\right)\right) \\ &= d_{\mathrm{TV}}\left(\mathcal{N}\left(\mu^*\cdot\frac{\alpha\sqrt{8\pi}}{\sqrt{d}},I\right),\mathcal{N}\left(m'\cdot\frac{\alpha\sqrt{8\pi}}{\sqrt{d}},I\right)\right) \\ &\ge \frac{c\cdot e^{-c^2/2}}{\sqrt{2\pi}} \qquad \text{if } \frac{1}{2}\|\mu^*-m'\|_2\frac{\alpha\sqrt{8\pi}}{\sqrt{d}} \ge c \\ &> 7\alpha \qquad \text{if } \|\mu^*-m'\|_2 \ge 30\frac{\sqrt{d}}{\sqrt{2\pi}}, \end{aligned}$$

where the final inequality follows by setting $c = 30\alpha \le$

1. Thus

$$\begin{aligned}
&|\{H \in \mathcal{C}_\alpha : d_{\mathrm{TV}}(H, \mathcal{N}(\mu, I)) \le 7\alpha\}| \\
&\le \left|\left\{m' \in \mathbb{Z}^d : \|\mu^* - m'\|_2 < 30\frac{\sqrt{d}}{\sqrt{2\pi}}\right\}\right| \\
&\le \left|\left\{m' \in \mathbb{Z}^d : \|m - m'\|_2 < 30\frac{\sqrt{d}}{\sqrt{2\pi}} + \|\mu^* - m'\|_2\right\}\right| \\
&\le \left|\left\{m' \in \mathbb{Z}^d : \|m - m'\|_2 < 13\sqrt{d}\right\}\right| \\
&\le \left|\left\{w \in \mathbb{Z}^d : \|w\|_1 < 13d\right\}\right| .
\end{aligned}$$

Now we note that any $w \in \mathbb{Z}^d$ with $\|w\|_1 \le r$ can be written as $w = x - y$ where $x, y \in \mathbb{Z}^d$ with $\sum_{i=1}^d x_i + y_i = r$ and, for all $i \in [d]$, we have $x_i \ge 0$ and $y_i \ge 0$. Instead of counting these w vectors, we can count such (x, y) vector pairs. We can interpret a pair of x, y vectors as a way of putting r balls into $2d$ bins or r "stars" and $2d - 1$ "bars". We can thus count $\left|\left\{w \in \mathbb{Z}^d : \|w\|_1 < 13d\right\}\right| \le \left|\left\{x, y \in \mathbb{Z}^d : \|x\|_1 + \|y\|_2 = 13d - 1, x \ge 0, y \ge 0\right\}\right| \le \binom{15d-2}{2d-1} \le 2^{15d}$. □

Applying Theorem IV.1 with the cover of Lemma VI.12 and the VC bound from Lemma VI.9 now yields an algorithm.

Corollary VI.13. *Suppose we are given a set of samples $X_1, \ldots, X_n \sim P$, where P is a spherical Gaussian distribution $\mathcal{N}(\mu, I)$ in d-dimensions. Then there exists a (ε, δ)-differentially private algorithm which outputs a spherical Gaussian distribution H^* such that $d_{\mathrm{TV}}(P, H^*) \le 7\alpha$ with probability $\ge 1 - 2^{-d}$, so long as*

$$n = \Omega\left(\frac{d}{\alpha^2} + \frac{d + \log(1/\delta)}{\alpha\varepsilon}\right).$$

Karwa and Vadhan [25] give an algorithm for estimating a univariate Gaussian with unbounded mean. One can consider applying their algorithm independently to the d coordinates (which is done in [17]), giving a sample complexity bound of $\tilde{O}\left(\frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{\sqrt{d}\log^{3/2}(1/\delta)}{\varepsilon}\right)$, which our bound dominates except for very small values of α.

2) Univariate Gaussians with Unbounded Mean and Variance: Our methods also allow us to derive learning algorithms for univariate Gaussians with unknown mean and variance.

Lemma VI.14. *For all α less then some absolute constant, there exists an α-cover $\mathcal{C}_\alpha$ of the set of univariate Gaussian distributions satisfying $\forall \mu, \sigma \in \mathbb{R}$ $\left|\left\{H \in \mathcal{C}_\alpha : d_{\mathrm{TV}}(H, \mathcal{N}(\mu, \sigma^2)) \le 7\alpha\right\}\right| \le O(1)$.*

Proof. For all $\mu, \tilde{\mu} \in \mathbb{R}$ and all $\sigma, \tilde{\sigma} > 0$, we have [104, Thm 1.3] $\frac{1}{200}\min\left\{1, \max\left\{\frac{|\tilde{\sigma}^2 - \sigma^2|}{\tilde{\sigma}^2}, \frac{40|\tilde{\mu} - \mu|}{\tilde{\sigma}}\right\}\right\} \le d_{\mathrm{TV}}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)) \le \frac{3|\tilde{\sigma}^2 - \sigma^2|}{2\tilde{\sigma}^2} + \frac{|\tilde{\mu} - \mu|}{2\tilde{\sigma}}$. Let $\beta = \alpha$ and $\gamma = \log(1 + \alpha/2)$. Define the set of distributions

$$\mathcal{C}_\alpha = \left\{\mathcal{N}\left(\beta e^{\gamma n} m, e^{2\gamma n}\right) : n, m \in \mathbb{Z}\right\}.$$

We first show that $\mathcal{C}_\alpha$ is an α-cover: Let $\mu \in \mathbb{R}$ and $\sigma > 0$. Let $n = \left[\frac{\log\sigma}{\gamma}\right]$ and $m = \left[\frac{\mu}{\beta e^{\gamma n}}\right]$, where $[x]$ denotes the nearest integer to x, satisfying $|x - [x]| \le \frac{1}{2}$. Let $\tilde{\sigma} = e^{\gamma n}$ and $\tilde{\mu} = \beta e^{\gamma n} m$ so that $e^{-\gamma} \le \frac{\tilde{\sigma}^2}{\sigma^2} \le e^{\gamma}$ and $|\mu - \tilde{\mu}| \le \frac{1}{2}\beta e^{\gamma n} = \frac{1}{2}\beta\tilde{\sigma}$. Thus $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2) \in \mathcal{C}_\alpha$ and $d_{\mathrm{TV}}(\mathcal{N}(\mu, \sigma^2), \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)) \le \frac{3}{2}\left(e^{\gamma} - 1\right) + \frac{\beta}{4} \le \alpha$, as required.

It only remains to show that the cover size is locally small. Let $\mu \in \mathbb{R}$ and $\sigma > 0$.

$$\begin{aligned}
&\left|\left\{H \in \mathcal{C}_\alpha : d_{\mathrm{TV}}(H, \mathcal{N}(\mu, \sigma^2)) \le 7\alpha\right\}\right| \\
&= \left|\left\{n, m \in \mathbb{Z} : d_{\mathrm{TV}}(\mathcal{N}\left(\beta e^{\gamma n} m, e^{2\gamma n}\right), \mathcal{N}(\mu, \sigma^2)) \le 7\alpha\right\}\right| \\
&\le \left|\left\{n, m \in \mathbb{Z} : \max\left\{\frac{|e^{2\gamma n} - \sigma^2|}{e^{2\gamma n}}, \frac{40|\beta e^{\gamma n} m - \mu|}{e^{\gamma n}}\right\} \le 1400\alpha\right\}\right| \\
&= \left|\left\{n, m \in \mathbb{Z} : \begin{array}{c} \frac{-\log(1+1400\alpha)}{2\gamma} \le n - \frac{\log\sigma}{\gamma} \le \frac{-\log(1-1400\alpha)}{2\gamma} \\ -35\frac{\alpha}{\beta} \le m - \frac{\mu}{\beta e^{\gamma n}} \le 35\frac{\alpha}{\beta} \end{array}\right\}\right| \\
&\le \left(\frac{-\log(1 - 1400\alpha)}{2\gamma} - \frac{-\log(1 + 1400\alpha)}{2\gamma} + 1\right) \cdot (35 - (-35) + 1) \\
&= \frac{1}{2\log(1 + \alpha/2)} \log\left(\frac{1 + 1400\alpha}{1 - 1400\alpha}\right) \cdot 71 + 71 \\
&= O(1).
\end{aligned}$$

□

Combining Lemma VI.14 with Lemma VI.9 and Theorem IV.1 yields the following.

Corollary VI.15. *Suppose we are given a set of samples $X_1, \ldots, X_n \sim P$, where P is a univariate Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Then there exists a (ε, δ)-differentially private algorithm which outputs a univariate Gaussian distribution H^* such that $d_{\mathrm{TV}}(P, H^*) \le 7\alpha$ with probability $\ge 9/10$, so long as*

$$n = \Omega\left(\frac{1}{\alpha^2} + \frac{\log(1/\delta)}{\alpha\varepsilon}\right).$$

This sample complexity is comparable to to that of Karwa and Vadhan [25], who give an (ε, δ)-DP algorithm with sample complexity $\tilde{O}\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}\right)$.

C. Sums of Independent Random Variables

In this section, we apply our results to distribution classes which are defined as the sum of independent (but not necessarily identical) distributions. These are all generalizations of the classical Binomial distribution, and they have enjoyed a great deal of study into the construction of sparse covers. To the best of our knowledge, we are the first to provide private learning algorithms for these classes.

We start with the Poisson Binomial distribution.

Definition VI.16. *A k-*Poisson Binomial Distribution *(k-PBD) is the sum of k independent Bernoulli random variables.*

We next consider sums of independent integer random variables, which generalize PBDs (which correspond to the case $d = 2$).

Definition VI.17. *A (k, d)-*Sum of Independent Integer Random Variables *((k, d)-SIIRV) is the sum of k independent random variables over $\{0, \ldots, d-1\}$.*

Finally, we consider Poisson Multinomial distributions, which again generalize PBDs (which, again, correspond to the case $d = 2$).

Definition VI.18. *A (k, d)-*Poisson Multinomial Distribution *((k, d)-PMD) is the sum of k independent d-dimensional categorical random variables, i.e., distributions over $\{e_1, \ldots, e_d\}$, where e_i is the ith basis vector.*

We start with a covering result for SIIRVs (including the special case of PBDs), which appears in [47]. Previous covers for PBDs and SIIRVs appear in [105], [106], [107].

Lemma VI.19 ([47])**.** *There exists an α-cover of the set of (k, d)-SIIRVs of size*

$$k \cdot 2^{O(d \log^2(1/\alpha) + d \log^2 d)}.$$

Using this cover, we can apply Corollary I.2 to attain the following learning result for PBDs and SIIRVs.

Corollary VI.20. *Suppose we are given a set of samples $X_1, \ldots, X_n \sim P$, where P is α-close to a (k, d)-SIIRV. Then for any constant $\zeta > 0$, there exists an ε-differentially private algorithm which outputs a (k, d)-SIIRV H^* such that $d_{\mathrm{TV}}(P, H^*) \leq (6 + 2\zeta)\alpha$ with probability $\geq 9/10$, so long as*

$$n = \Omega\left(\left(\log k + d \log^2(1/\alpha) + d \log^2 d\right)\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)\right).$$

Next, we move on to PMDs. The following cover does not appear verbatim in any single location, but is a combination of results from a few different sources. The proofs for the best bounds on first term appears in [46], the second in [45], and the third in [49]. Larger covers previously appeared in [108], [109].

Lemma VI.21 ([45], [46], [49])**.** *For any $d > 2$, there exists an α-cover of the set of (k, d)-PMDs of size*

$$k^{O(d)} \cdot \min\left\{2^{\mathrm{poly}(d/\alpha)}, (1/\alpha)^{O(d \log(d/\alpha)/\log\log(d/\alpha))^{d-1}}\right\}.$$

This implies the following learning result for PMDs.

Corollary VI.22. *Suppose we are given a set of samples $X_1, \ldots, X_n \sim P$, where P is α-close to a (k, d)-PMD, for any $d > 2$. Then there exists an ε-differentially private algorithm which outputs a (k, d)-PMD H^* such that $d_{\mathrm{TV}}(P, H^*) \leq (6 + 2\zeta)\alpha$ with probability $\geq 9/10$, so long as $n = \tilde{\Omega}((d \log k + \min\{\mathrm{poly}(\frac{d}{\alpha}), O(\frac{d \log(d/\alpha)}{\log\log(d/\alpha)})^{d-1} \cdot \log(1/\alpha)\})(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}))$.*

D. Piecewise Polynomials

In this section, we apply our results to semi-agnostically learn piecewise polynomials. This class of distributions is very expressive, allowing us to approximate a wide range of natural distribution classes.

Definition VI.23. *A (t, d, k)-*piecewise polynomial distribution *is a distribution P over $[k]$, such that there exists a partition of $[k]$ into t disjoint intervals $I_1, \ldots, I_t$ such that on each interval $I_j \subseteq [k]$, the probability mass function of P takes the form $p_j(x) = \sum_{i=0}^{d} c_i^{(j)} x^i$ for some coefficients $c_i^{(j)}$, for all $x \in I_j$.*

We construct a cover for piecewise polynomials.

Lemma VI.24. *There exists a universal constant $c > 0$ such that there is an α-cover of the set of (t, d, k)-piecewise polynomials of size*

$$\binom{k}{t-1} \cdot \left(\frac{tk \cdot e^{cd^{1/2}}}{\alpha}\right)^{(d+1)t}.$$

Proof. We specify an element of the cover by

1) Selecting one of $\binom{k}{t-1}$ partitions of $[k]$ into t intervals $I_1, \ldots, I_t$, and
2) For each interval I_j, selecting an element of an (α/t)-cover $\mathcal{C}_j$ of the set of degree-d polynomials over I_j which are uniformly bounded by 1.

The total size of the cover is $\binom{k}{t-1} \prod_{j=1}^{t} |\mathcal{C}_j|$. The theorem follows from Proposition VI.25 below, which constructs an (α/t)-cover $\mathcal{C}_j$ of size at most $\left(\frac{tk \cdot e^{cd^{1/2}}}{\alpha}\right)^{d+1}$ for every interval I_j.

□

Proposition VI.25. *There exist constants $b, c > 0$ for which the following holds. Let $I \subseteq [k]$ be an interval and let $\mathcal{P}$ be the set of polynomials $p : I \to \mathbb{R}$ of degree d such that $|p(x)| \leq 1$ for all $x \in I$. There exists an α-cover of $\mathcal{P}$ of size*

$$\min\left\{\left(\frac{2k}{\alpha}\right)^{|I|}, \left(\frac{ckd^2 \cdot e^{bd^2/|I|}}{\alpha}\right)^{d+1}\right\}.$$

The proof of Proposition VI.25 relies on two major results in approximation theory, which we now state.

Lemma VI.26 (Duffin and Schaeffer [110])**.** *Let $p : [-1, 1] \to \mathbb{R}$ be a polynomial such that $|p(x)| \leq 1$ for* *all x of the form $x = \cos(j\pi/d)$ for $j = 0, 1, \dots, d$. Then $|p'(x)| \le d^2$ for all $x \in [-1, 1]$.*

Lemma VI.27 (Coppersmith and Rivlin [111])**.** *There exist constants $a, b > 0$ for which the following holds. Let $p : \mathbb{R} \to \mathbb{R}$ be a polynomial of degree d, and suppose that $|p(t)| \le 1$ for all $t = 0, 1, \dots, m$. Then $|p(t)| \le a\exp(bd^2/m)$ for all $t \in [0, m]$.*

Proof of Proposition VI.25. We consider two cases, corresponding to the two terms in the minimum. First, consider the function $f : I \to \mathbb{R}$ where $f(t)$ is obtained by rounding $p(t)$ to the nearest multiple of α/k. Then f satisfies $\sum_{t\in I} |f(t) - p(t)| \le \alpha$. There are at most $(2k/\alpha)^{|I|}$ functions f which can be constructed this way, giving the first term in the maximum.

For the second term, we construct a cover for $\mathcal{P}$ by approximately interpolating through $d + 1$ carefully chosen points in the *continuous* interval corresponding to I. By applying an affine shift, we may assume that $I = \{0, 1, \dots, m\}$ for some integer $m \le k - 1$. Let $p \in \mathcal{P}$ and for $x \in [0, m]$ let $\hat{p}(x)$ be the value of $p(x)$ rounded to the nearest integer multiple of $\alpha/(2kd^2)$. Let $q : [0, m] \to \mathbb{R}$ be the unique degree-d polynomial obtained by interpolating through the points $(x_j, \hat{p}(x_j))$ where $x_j = m(1 + \cos(j\pi/d))/2$ for $j = 0, 1, \dots, d$.

We first argue that the polynomial q so defined satisfies $\sum_{t=0}^{m} |p(t) - q(t)| \le \alpha$. Let $r(x) = p(x) - q(x)$ for $x \in [0, m]$. Then by construction, $|r(x_j)| \le \alpha/(2kd^2)$ for all interpolation points x_j. By the Duffin-Schaeffer Inequality (Lemma VI.26), we therefore have $|r'(x)| \le \frac{\alpha}{km}$ for all $x \in [0, m]$. By the Fundamental Theorem of Calculus, $r(t) = r(0) + \int_0^t r'(t)\,dt$ satisfies $|r(t)| \le (t + 1) \cdot \frac{\alpha}{km} \le \alpha/k$, and hence $\sum_{t=0}^{m} |r(t)| \le \alpha$.

We now argue that the set of polynomials q that can be constructed in this fashion has size $(ckd^2 \exp(bd^2/m)/\alpha)^{d+1}$. By the Coppersmith-Rivlin Inequality (Lemma VI.27), there are constants $a, b > 0$ such that $|p(x)| \le a\exp(bd^2/m)$ for all $x \in [0, m]$. Therefore, for each $p \in \mathcal{P}$ and each interpolation point x_j, there are at most $4a \cdot kd^2 \exp(bd^2/m)/\alpha$ possible values that $\hat{p}(x_j)$ can take. Hence, the polynomial q can take one of at most $(4a \cdot kd^2 \exp(bd^2/m)/\alpha)^{d+1}$ possible values, as we wanted to show. □

Lemma VI.28. *The VC dimension of (t, d, k)-piecewise polynomial distributions is at most $2t(d + 1)$.*

Proof. Consider two piecewise polynomial distributions. The difference between their probability mass functions is a piecewise polynomial of degree $\le d$. The number of intervals needed to represent this piecewise function is $\le 2t$. It follows that this difference can change sign at most $2td + 2t - 1$ times – each polynomial can change sign at most d times and the sign can change at the interval boundaries. Thus such a function cannot label $2td + 2t + 1$ points with alternating signs, which implies the VC bound. □

As a corollary, we obtain the following learning algorithm.

Corollary VI.29. *Suppose we are given a set of samples $X_1, \dots, X_n \sim P$, where P is α-close to a (t, d, k)-piecewise polynomial. Then there exists an ε-differentially private algorithm which outputs a (t, d, k)-piecewise polynomial H^* such that $d_{\mathrm{TV}}(P, H^*) \le (6 + 2\zeta)\alpha$ with probability $\ge 9/10$, so long as $n = \Omega\left(\frac{(d+1)t}{\alpha^2} + \frac{(d+1)t}{\alpha\varepsilon} \cdot \left(\sqrt{d+1}\log k + \log\left(\frac{t}{\alpha}\right)\right)\right)$.*

We compare with the work of Diakonikolas, Hardt, and Schmidt [18]. They present an efficient algorithm for $(t, 1, k)$-piecewise polynomials, with sample complexity $\tilde{O}\left(\frac{t}{\alpha^2} + \frac{t\log k}{\alpha\varepsilon}\right)$, which our algorithm matches.[7] They also claim their results extend to (t, d, k)-piecewise polynomials, though no theorem statement is provided. While we have not investigated the details of this extension, we believe the resulting sample complexity should be qualitatively similar to ours, plausibly with the factor of $\frac{t(d+1)^{3/2}\log k}{\alpha\varepsilon}$ replaced by $\frac{t(d+1)\log k}{\alpha\varepsilon}$.

E. Mixtures

In this section, we show that our results immediately extend to learning mixtures of classes of distributions.

Definition VI.30. *Let $\mathcal{H}$ be some set of distributions. A k-mixture of $\mathcal{H}$ is a distribution with density $\sum_{i=1}^{k} w_i P_i$, where each $P_i \in \mathcal{H}$.*

Our results follow roughly due to the fact that a cover for k-mixtures of a class can be written as the Cartesian product of k covers for the class. More precisely, we state the following result which bounds the size of the cover of the set of k-mixtures.

Lemma VI.31. *Consider the class of k-mixtures of $\mathcal{H}$, where $\mathcal{H}$ is some set of distributions. There exists a 2α-cover of this class of size $|\mathcal{C}_\alpha|^k \left(\frac{k}{2\alpha} + 1\right)^{k-1}$, where $\mathcal{C}_\alpha$ is an α-cover of $\mathcal{H}$.*

Proof. Each element in the cover of the class of mixtures will be obtained by taking k distributions from $\mathcal{C}_\alpha$, in combination with k mixing weights, which are selected from the set $\left\{0, \frac{2\alpha}{k}, \frac{4\alpha}{k}, \dots, 1\right\}$, such that the sum of the mixing weights is 1. The size of this cover is $|\mathcal{C}_\alpha|^k \cdot \left(\frac{k}{2\alpha} + 1\right)^{k-1}$. We reason about the accuracy of the cover as follows. Fix some mixture of k distributions as $\sum_{i=1}^{k} w_i^{(1)} P_i^{(1)}$, and we will reason about the closest element in

[7] As stated in [18], their algorithm guarantees approximate differential privacy, but swapping in an appropriate pure DP subroutine gives this result.

our cover, $\sum_{i=1}^k w_i^{(2)} P_i^{(2)}$. By triangle inequality, we have that $d_{\mathrm{TV}}\left(\sum_{i=1}^k w_i^{(1)} P_i^{(1)}, \sum_{i=1}^k w_i^{(2)} P_i^{(2)}\right) \leq \sum_{i=1}^k \frac{1}{2}\left|w_i^{(1)} - w_i^{(2)}\right| + w_i^{(1)} d_{\mathrm{TV}}\left(P_i^{(1)}, P_i^{(2)}\right)$. Since $\mathcal{C}_\alpha$ is an α-cover and $\sum_{i=1}^k w_i^{(1)} = 1$, the total variation distance incurred by the second term will be at most α. As for the mixing weights, note that for the first $k-1$ weights, the nearest weight is at distance at most $\frac{\alpha}{k}$, contributing a total of less than $\frac{\alpha}{2}$. The last mixing weight can be rewritten in terms of the sum of the errors of the other mixing weights, similarly contributing another total of less than $\frac{\alpha}{2}$. This results in the total error being at most 2α, as desired. □

With this in hand, the following corollary is almost immediate from Corollary I.2. The factor of $(9+3\zeta)\alpha$ (as opposed to $(6+2\zeta)\alpha$) is because the closest distribution in the cover of mixture distributions is 3α-close to be P (rather than 2α).

Corollary VI.32. *Let $X_1, \ldots, X_n \sim P$, where P is α-close to a k-mixture of distributions from some set $\mathcal{H}$. Let $\mathcal{C}_\alpha$ be an α-cover of the set $\mathcal{H}$, and $\zeta > 0$ be a constant. There exists an ε-differentially private algorithm which outputs a distribution which is $(9+3\zeta)\alpha$-close to P with probability $\geq 9/10$, as long as*

$$n = \Omega\left((k \log |\mathcal{C}_\alpha| + k \log(k/\alpha))\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)\right).$$

For example, instantiating this for mixtures of Gaussians (and disregarding terms which depend on R and κ), we get an algorithm with sample complexity $\tilde{O}\left(\frac{kd^2}{\alpha^2} + \frac{kd^2}{\alpha\varepsilon}\right)$.

F. Supervised Learning

We describe an application of our results to the task of binary classification, as modeled by differentially private PAC learning [88]. Let $\mathcal{F} = \{f : X \to \{0,1\}\}$ be a publicly known *concept class* of Boolean functions over a domain X. Let P be an unknown probability distribution over X, and let f be an unknown function from $\mathcal{F}$. Given a sequence $\{(x_i, f(x_i))\}_{i=1}^n$ of i.i.d. samples from P together with their labels under f, the goal of a PAC learner L is to identify a hypothesis $h : X \to \{0,1\}$ such that $\Pr_{x\sim P}[h(x) \neq f(x)] \leq \alpha$ for some error parameter $\alpha > 0$. We say that L is (α,β)-*accurate* if for every $f \in \mathcal{F}$ and every distribution P, it is able to identify such a hypothesis h with probability at least $1-\beta$ over the choice of the sample and any internal randomness of L.

One of the core results of statistical learning theory is that the sample complexity of *non-private* PAC learning is characterized, up to constant factors, by the VC dimension of the concept class $\mathcal{F}$. When one additionally requires the learner L to be differentially private with respect to its input sample, such a characterization is unknown. However, it is known that the sample complexity of private learning can be arbitrarily higher than that of non-private learning. For example, when $\mathcal{F} = \{f_t : t \in X\}$ is the class of threshold functions defined by $f_t(x) = 1 \iff x \leq t$ over a totally ordered domain X, the sample complexity of PAC learning under the most permissive notion of (ε,δ)-differential privacy is $\Omega(\log^* |X|)$ [29], [112]. Meanwhile, the VC dimension of this class, and hence the sample complexity of non-private learning, is a constant independent of $|X|$.

While this separation shows that there can be a sample cost of privacy for PAC learning, this cost can be completely eliminated if the distribution P on examples is known. This was observed by Beimel, Nissim, and Stemmer [113], who showed that if a good approximation to P is known, e.g., from public unlabeled examples or from differentially private processing of unlabeled examples, then the number of labeled examples needed for private PAC learning is only $O(VC(\mathcal{F}))$.

Theorem VI.33. *Let $\varepsilon > 0$, $\mathcal{F} = \{f : X \to \{0,1\}\}$, and P be a publicly known distribution over X. For $n = O\left(\frac{1}{\alpha^2\varepsilon}(VC(\mathcal{F})\log(1/\alpha) + \log(1/\beta))\right)$, there exists an ε-differentially private algorithm $L : (X \times \{0,1\})^n \to \mathcal{F}$ such that for every $f \in \mathcal{F}$, with probability at least $1-\beta$ over the choice of $x_1, \ldots, x_n \leftarrow P$, we have that $L((x_1, f(x_1)), \ldots, (x_n, f(x_n)))$ produces $h \in \mathcal{F}$ such that $\Pr_{x\sim P}[f(x) \neq h(x)] \leq \alpha$.*

Our results suggest a natural two-step algorithm for private PAC learning when the distribution P itself is not known, but is known to (approximately) come from a set of distributions $\mathcal{H}$: The algorithm first uses private hypothesis selection to select $\hat{H}$ with $d_{\mathrm{TV}}(P, \hat{H}) \leq \alpha/2$, and then runs the algorithm of [113] using $\hat{H}$ in place of P with error parameter $\alpha/2$. Using the fact that $d_{\mathrm{TV}}(P, \hat{H}) \leq \alpha/2$ implies $|\Pr_{x\sim P}[f(x) \neq h(x)] - \Pr_{x\sim \hat{H}}[f(x) \neq h(x)]| \leq \alpha/2$, the following result holds by combining Theorem VI.33 with Corollary I.2.

Corollary VI.34. *Let $\mathcal{H}$ be a set of distributions over X with an α-cover $\mathcal{C}_\alpha$. Let P be a distribution over X with $d_{\mathrm{TV}}(P, \mathcal{H}) \leq \alpha/(4(3+\zeta))$. Then for*

$$n = O\left(\frac{\log |\mathcal{C}_\alpha|}{\alpha^2} + \frac{\log |\mathcal{C}_\alpha|}{\alpha\varepsilon} + \frac{VC(\mathcal{F})\log(1/\alpha)}{\alpha^2\varepsilon}\right)$$

there exists an ε-differentially private algorithm $L : (X \times \{0,1\})^n \to \mathcal{F}$ such that for every $f \in \mathcal{F}$, with probability at least $3/4$ over the choice of $x_1, \ldots, x_n \leftarrow P$, we have that $L((x_1, f(x_1)), \ldots, (x_n, f(x_n)))$ produces $h \in \mathcal{F}$ such that $\Pr_{x\sim P}[f(x) \neq h(x)] \leq \alpha$.

Theorem VI.33 can, of course, also be combined with the more refined guarantees of Theorem IV.1.

As an example application, combining Theorem VI.33 with Corollary VI.13 gives a (ε, δ)-differentially private algorithm for learning one-dimension thresholds with respect to univariate Gaussian distributions on the reals. In contrast, this task is impossible without making distributional assumptions.

VII. Conclusions

In this paper, we presented differentially private methods for hypothesis selection. The sample complexity can be bounded by the logarithm of the number of hypotheses. This allows us to provide bounds on the sample complexity of (semi-agnostically) learning a class which depend on the logarithm of the covering number, complementing known lower bounds which depend on the logarithm of the packing number. There are many interesting questions left open by our work, a few of which we outline below.

1) Our algorithms for learning classes of distributions all use cover-based arguments, and thus are not computationally efficient. For instance, we provide the first $\tilde{O}(d)$ sample complexity upper bound on ε-differentially privately learning a product distribution and Gaussian with known covariance. One interesting question is whether there is an efficient algorithm which achieves this sample complexity.
2) The running time of our method is quadratic in the number of hypotheses – is it possible to reduce this to a near-linear time complexity?
3) Our main theorem obtains an approximation factor which is arbitrarily close to 3, which is optimal for this problem, even without privacy. This factor can be reduced to 2 if one is OK with outputting a mixture of hypotheses from the set [12]. Is this achievable with privacy constraints?

Acknowledgment

The authors would like to thank Shay Moran for bringing to their attention the application to PAC learning mentioned in Section VI-F, Jonathan Ullman for asking questions which motivated Remark VI.4, Clément Canonne for assistance in reducing the constant factor in the approximation guarantee, and the anonymous reviewers for many helpful and encouraging comments and suggestions.

References

[1] Y. G. Yatracos, "Rates of convergence of minimum distance estimators and Kolmogorov's entropy," *The Annals of Statistics*, vol. 13, no. 2, pp. 768–774, 1985.

[2] L. Devroye and G. Lugosi, "A universally acceptable smoothing factor for kernel density estimation," *The Annals of Statistics*, vol. 24, no. 6, pp. 2499–2512, 1996.

[3] ——, "Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes," *The Annals of Statistics*, vol. 25, no. 6, pp. 2626–2637, 1997.

[4] ——, *Combinatorial methods in density estimation*. Springer, 2001.

[5] S. Mahalanabis and D. Stefankovic, "Density estimation in linear time," in *Proceedings of the 21st Annual Conference on Learning Theory*, ser. COLT '08, 2008, pp. 503–512.

[6] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning Poisson binomial distributions," in *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing*, ser. STOC '12. New York, NY, USA: ACM, 2012, pp. 709–728.

[7] C. Daskalakis and G. Kamath, "Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians," in *Proceedings of the 27th Annual Conference on Learning Theory*, ser. COLT '14, 2014, pp. 1183–1213.

[8] A. T. Suresh, A. Orlitsky, J. Acharya, and A. Jafarpour, "Near-optimal-sample estimators for spherical Gaussian mixtures," in *Advances in Neural Information Processing Systems 27*, ser. NIPS '14. Curran Associates, Inc., 2014, pp. 1395–1403.

[9] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Sorting with adversarial comparators and application to density estimation," in *Proceedings of the 2014 IEEE International Symposium on Information Theory*, ser. ISIT '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1682–1686.

[10] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart, "Robust estimators in high dimensions without the computational intractability," in *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '16. Washington, DC, USA: IEEE Computer Society, 2016, pp. 655–664.

[11] J. Acharya, M. Falahatgar, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Maximum selection and sorting with adversarial comparators," *Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2427–2457, 2018.

[12] O. Bousquet, D. M. Kane, and S. Moran, "The optimal approximation factor in density estimation," in *Proceedings of the 32nd Annual Conference on Learning Theory*, ser. COLT '19, 2019, pp. 318–341.

[13] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the 3rd Conference on Theory of Cryptography*, ser. TCC '06. Berlin, Heidelberg: Springer, 2006, pp. 265–284.

[14] Differential Privacy Team, Apple, "Learning with privacy at scale," https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf, December 2017.

[15] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, ser. CCS '14. New York, NY, USA: ACM, 2014, pp. 1054–1067.

[16] A. N. Dajani, A. D. Lauger, P. E. Singer, D. Kifer, J. P. Reiter, A. Machanavajjhala, S. L. Garfinkel, S. A. Dahl, M. Graham, V. Karwa, H. Kim, P. Lelerc, I. M. Schmutte, W. N. Sexton, L. Vilhuber, and J. M. Abowd, "The modernization of statistical disclosure limitation at the U.S. census bureau," 2017, presented at the September 2017 meeting of the Census Scientific Advisory Committee.

[17] G. Kamath, J. Li, V. Singhal, and J. Ullman, "Privately learning high-dimensional distributions," in *Proceedings of the 32nd Annual Conference on Learning Theory*, ser. COLT '19, 2019, pp. 1853–1902.

[18] I. Diakonikolas, M. Hardt, and L. Schmidt, "Differentially private learning of structured discrete distributions," in *Advances in Neural Information Processing Systems 28*, ser. NIPS '15. Curran Associates, Inc., 2015, pp. 2566–2574.

[19] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt, "Sample-optimal density estimation in nearly-linear time," in *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algo-*

rithms, ser. SODA '17. Philadelphia, PA, USA: SIAM, 2017, pp. 1278–1289.

[20] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 94–103.

[21] C. Dwork and J. Lei, "Differential privacy and robust statistics," in *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, ser. STOC '09. New York, NY, USA: ACM, 2009, pp. 371–380.

[22] A. G. Thakurta and A. Smith, "Differentially private feature selection via stability arguments, and the robustness of the lasso," in *Proceedings of the 26th Annual Conference on Learning Theory*, ser. COLT '13, 2013, pp. 819–850.

[23] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke, "Composable and versatile privacy via truncated cdp," in *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, ser. STOC '18. New York, NY, USA: ACM, 2018, pp. 74–86.

[24] C. L. Canonne, G. Kamath, A. McMillan, A. Smith, and J. Ullman, "The structure of optimal private tests for simple hypotheses," in *Proceedings of the 51st Annual ACM Symposium on the Theory of Computing*, ser. STOC '19. New York, NY, USA: ACM, 2019, pp. 310–321.

[25] V. Karwa and S. Vadhan, "Finite sample differentially private confidence intervals," in *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ser. ITCS '18. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, pp. 44:1–44:9.

[26] T. T. Cai, Y. Wang, and L. Zhang, "The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy," *arXiv preprint arXiv:1902.04495*, 2019.

[27] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, ser. STOC '07. New York, NY, USA: ACM, 2007, pp. 75–84.

[28] G. Kamath, O. Sheffet, V. Singhal, and J. Ullman, "Differentially private algorithms for learning mixtures of separated Gaussians," in *Advances in Neural Information Processing Systems 32*, ser. NeurIPS '19. Curran Associates, Inc., 2019, pp. 168–180.

[29] M. Bun, K. Nissim, U. Stemmer, and S. Vadhan, "Differentially private release and learning of threshold functions," in *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 634–649.

[30] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The SuLQ framework," in *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '05. New York, NY, USA: ACM, 2005, pp. 128–138.

[31] M. Bun, J. Ullman, and S. Vadhan, "Fingerprinting codes and the price of approximate differential privacy," in *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, ser. STOC '14. New York, NY, USA: ACM, 2014, pp. 1–10.

[32] T. Steinke and J. Ullman, "Between pure and approximate differential privacy," *The Journal of Privacy and Confidentiality*, vol. 7, no. 2, pp. 3–22, 2017.

[33] J. Acharya, G. Kamath, Z. Sun, and H. Zhang, "Inspectre: Privately estimating the unseen," in *Proceedings of the 35th International Conference on Machine Learning*, ser. ICML '18. JMLR, Inc., 2018, pp. 30–39.

[34] A. Smith, "Privacy-preserving statistical estimation with optimal convergence rates," in *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, ser. STOC '11. New York, NY, USA: ACM, 2011, pp. 813–822.

[35] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 429–438.

[36] S. Wang, L. Huang, P. Wang, Y. Nie, H. Xu, W. Yang, X.-Y. Li, and C. Qiao, "Mutual information optimally local private discrete distribution estimation," *arXiv preprint arXiv:1607.08025*, 2016.

[37] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proceedings of the 33rd International Conference on Machine Learning*, ser. ICML '16. JMLR, Inc., 2016, pp. 2436–2444.

[38] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: Estimating distributions privately, efficiently, and with little communication," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, ser. AISTATS '19. JMLR, Inc., 2019, pp. 1120–1129.

[39] J. C. Duchi and F. Ruan, "The right complexity measure in locally private estimation: It is not the Fisher information," *arXiv preprint arXiv:1806.05756*, 2018.

[40] M. Joseph, J. Kulkarni, J. Mao, and Z. S. Wu, "Locally private Gaussian estimation," in *Advances in Neural Information Processing Systems 32*, ser. NeurIPS '19. Curran Associates, Inc., 2019, pp. 2980–2989.

[41] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under locally differential privacy," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5662–5676, 2018.

[42] M. Gaboardi, R. Rogers, and O. Sheffet, "Locally private confidence intervals: Z-test and tight confidence intervals," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, ser. AISTATS '19. JMLR, Inc., 2019, pp. 2545–2554.

[43] G. Kamath and J. Ullman, "A primer on private statistics," *arXiv preprint arXiv:2005.00010*, 2020.

[44] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie, "On the learnability of discrete distributions," in *Proceedings of the 26th Annual ACM Symposium on the Theory of Computing*, ser. STOC '94. New York, NY, USA: ACM, 1994, pp. 273–282.

[45] C. Daskalakis, G. Kamath, and C. Tzamos, "On the structure, covering, and learning of Poisson multinomial distributions," in *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1203–1217.

[46] C. Daskalakis, A. De, G. Kamath, and C. Tzamos, "A size-free CLT for Poisson multinomials and its applications," in *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, ser. STOC '16. New York, NY, USA: ACM, 2016, pp. 1074–1086.

[47] I. Diakonikolas, D. M. Kane, and A. Stewart, "Optimal learning via the Fourier transform for sums of independent integer random variables," in *Proceedings of the 29th Annual Conference on Learning Theory*, ser. COLT '16, 2016, pp. 831–849.

[48] ——, "Properly learning Poisson binomial distributions in almost polynomial time," in *Proceedings of the 29th Annual Conference on Learning Theory*, ser. COLT '16, 2016, pp. 850–878.

[49] ——, "The Fourier transform of Poisson multinomial distributions and its algorithmic applications," in *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, ser. STOC '16. New York, NY, USA: ACM, 2016, pp. 1060–1073.

[50] A. De, P. M. Long, and R. A. Servedio, "Learning sums of independent random variables with sparse collective support," in *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '18. Washington, DC, USA: IEEE Computer Society, 2018, pp. 297–308.

[51] C. Daskalakis, I. Diakonikolas, and R. A. Servedio, "Learning k-modal distributions via testing," in *Proceedings of the 23th Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '12. Philadelphia, PA, USA: SIAM, 2012, pp. 1371–1385.

[52] S. O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun, "Efficient density estimation via piecewise polynomial approximation," in *Proceedings of the 46th Annual ACM Symposium*

on the Theory of Computing, ser. STOC '14. New York, NY, USA: ACM, 2014, pp. 604–613.

[53] ——, "Near-optimal density estimation in near-linear time using variable-width histograms," in *Advances in Neural Information Processing Systems 27*, ser. NIPS '14. Curran Associates, Inc., 2014, pp. 1844–1852.

[54] J. Acharya, C. Daskalakis, and G. Kamath, "Optimal testing for properties of distributions," in *Advances in Neural Information Processing Systems 28*, ser. NIPS '15. Curran Associates, Inc., 2015, pp. 3577–3598.

[55] S. Dasgupta, "Learning mixtures of Gaussians," in *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 634–644.

[56] S. Dasgupta and L. J. Schulman, "A two-round variant of EM for Gaussian mixtures," in *Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence*, ser. UAI '00. Morgan Kaufmann, 2000, pp. 152–159.

[57] S. Arora and R. Kannan, "Learning mixtures of arbitrary Gaussians," in *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, ser. STOC '01. New York, NY, USA: ACM, 2001, pp. 247–257.

[58] S. Vempala and G. Wang, "A spectral algorithm for learning mixtures of distributions," in *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 113–123.

[59] D. Achlioptas and F. McSherry, "On spectral learning of mixtures of distributions," in *Proceedings of the 18th Annual Conference on Learning Theory*, ser. COLT '05. Springer, 2005, pp. 458–469.

[60] K. Chaudhuri and S. Rao, "Learning mixtures of product distributions using correlations and independence," in *Proceedings of the 21st Annual Conference on Learning Theory*, ser. COLT '08, 2008, pp. 9–20.

[61] ——, "Beyond Gaussians: Spectral methods for learning mixtures of heavy-tailed distributions," in *Proceedings of the 21st Annual Conference on Learning Theory*, ser. COLT '08, 2008, pp. 21–32.

[62] A. Kumar and R. Kannan, "Clustering with spectral norm and the k-means algorithm," in *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 299–308.

[63] P. Awasthi and O. Sheffet, "Improved spectral-norm bounds for clustering," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, ser. APPROX '12. Springer, 2012, pp. 37–49.

[64] O. Regev and A. Vijayaraghavan, "On learning mixtures of well-separated Gaussians," in *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '17. Washington, DC, USA: IEEE Computer Society, 2017, pp. 85–96.

[65] S. B. Hopkins and J. Li, "Mixture models, robustness, and sum of squares proofs," in *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, ser. STOC '18. New York, NY, USA: ACM, 2018, pp. 1021–1034.

[66] I. Diakonikolas, D. M. Kane, and A. Stewart, "List-decodable robust mean estimation and learning mixtures of spherical Gaussians," in *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, ser. STOC '18. New York, NY, USA: ACM, 2018, pp. 1047–1060.

[67] P. Kothari, J. Steinhardt, and D. Steurer, "Robust moment estimation and improved clustering via sum of squares," in *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, ser. STOC '18. New York, NY, USA: ACM, 2018, pp. 1035–1046.

[68] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two Gaussians," in *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*, ser. STOC '10. New York, NY, USA: ACM, 2010, pp. 553–562.

[69] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 93–102.

[70] M. Belkin and K. Sinha, "Polynomial learning of distribution families," in *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 103–112.

[71] D. Hsu and S. M. Kakade, "Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions," in *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ser. ITCS '13. New York, NY, USA: ACM, 2013, pp. 11–20.

[72] J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. R. Voss, "The more, the merrier: the blessing of dimensionality for learning large Gaussian mixtures," in *Proceedings of the 27th Annual Conference on Learning Theory*, ser. COLT '14, 2014, pp. 1135–1164.

[73] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan, "Smoothed analysis of tensor decompositions," in *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, ser. STOC '14. New York, NY, USA: ACM, 2014, pp. 594–603.

[74] M. Hardt and E. Price, "Sharp bounds for learning a mixture of two Gaussians," in *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, ser. STOC '15. New York, NY, USA: ACM, 2015, pp. 753–760.

[75] R. Ge, Q. Huang, and S. M. Kakade, "Learning mixtures of Gaussians in high dimensions," in *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, ser. STOC '15. New York, NY, USA: ACM, 2015, pp. 761–770.

[76] J. Xu, D. J. Hsu, and A. Maleki, "Global analysis of expectation maximization for mixtures of two Gaussians," in *Advances in Neural Information Processing Systems 29*, ser. NIPS '16. Curran Associates, Inc., 2016, pp. 2676–2684.

[77] C. Daskalakis, C. Tzamos, and M. Zampetakis, "Ten steps of EM suffice for mixtures of two Gaussians," in *Proceedings of the 30th Annual Conference on Learning Theory*, ser. COLT '17, 2017, pp. 704–710.

[78] H. Ashtiani, S. Ben-David, N. Harvey, C. Liaw, A. Mehrabian, and Y. Plan, "Nearly tight sample complexity bounds for learning mixtures of Gaussians via sample compression schemes," in *Advances in Neural Information Processing Systems 31*, ser. NeurIPS '18. Curran Associates, Inc., 2018, pp. 3412–3421.

[79] J. Feldman, R. O'Donnell, and R. A. Servedio, "PAC learning axis-aligned mixtures of Gaussians with no separation assumption," in *Proceedings of the 19th Annual Conference on Learning Theory*, ser. COLT '06. Berlin, Heidelberg: Springer, 2006, pp. 20–34.

[80] ——, "Learning mixtures of product distributions over discrete domains," *SIAM Journal on Computing*, vol. 37, no. 5, pp. 1536–1564, 2008.

[81] J. Li and L. Schmidt, "Robust proper learning for mixtures of Gaussians via systems of polynomial inequalities," in *Proceedings of the 30th Annual Conference on Learning Theory*, ser. COLT '17, 2017, pp. 1302–1382.

[82] I. Aden-Ali, H. Ashtiani, and G. Kamath, "On the sample complexity of privately learning unbounded high-dimensional gaussians," in *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ser. ALT '21. JMLR, Inc., 2021.

[83] G. Kamath, V. Singhal, and J. Ullman, "Private mean estimation of heavy-tailed distributions," in *Proceedings of the 33rd Annual Conference on Learning Theory*, ser. COLT '20, 2020, pp. 2204–2235.

[84] Y. Liu, A. T. Suresh, F. Yu, S. Kumar, and M. Riley, "Learning discrete distributions: User vs item-level privacy," in *Advances in Neural Information Processing Systems 33*, ser. NeurIPS '20. Curran Associates, Inc., 2020.

[85] S. Gopi, G. Kamath, J. Kulkarni, A. Nikolov, Z. S. Wu, and H. Zhang, "Locally private hypothesis selection," in *Proceedings of the 33rd Annual Conference on Learning Theory*, ser. COLT '20, 2020, pp. 1785–1816.

[86] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[87] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '03. New York, NY, USA: ACM, 2003, pp. 211–222.

[88] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, 2011.

[89] C. Dwork, M. Naor, T. Pitassi, G. N. Rothblum, and S. Yekhanin, "Pan-private streaming algorithms," in *Proceedings of the 1st Conference on Innovations in Computer Science*, ser. ICS '10. Beijing, China: Tsinghua University Press, 2010, pp. 66–80.

[90] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev, "Distributed differential privacy via shuffling," in *Proceedings of the 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, ser. EUROCRYPT '19. Berlin, Heidelberg: Springer, 2019, pp. 375–403.

[91] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '19. Philadelphia, PA, USA: SIAM, 2019, pp. 2468–2479.

[92] A. Cheu and J. Ullman, "The limits of pan privacy and shuffle privacy for learning and estimation," *arXiv preprint arXiv:2009.08000*, 2020.

[93] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *arXiv preprint arXiv:1603.01887*, 2016.

[94] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Proceedings of the 14th Conference on Theory of Cryptography*, ser. TCC '16-B. Berlin, Heidelberg: Springer, 2016, pp. 635–658.

[95] I. Mironov, "Rényi differential privacy," in *Proceedings of the 30th IEEE Computer Security Foundations Symposium*, ser. CSF '17. Washington, DC, USA: IEEE Computer Society, 2017, pp. 263–275.

[96] V. Vapnik and A. Chervonenkis, *Theory of Pattern Recognition*. Nauka, 1974.

[97] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *The Annals of Probability*, vol. 22, no. 1, pp. 28–76, 1994.

[98] M. Hardt and K. Talwar, "On the geometry of differential privacy," in *Proceedings of the 42nd Annual ACM Symposium on the Theory of Computing*, ser. STOC '10. New York, NY, USA: ACM, 2010, pp. 705–714.

[99] A. Beimel, H. Brenner, S. P. Kasiviswanathan, and K. Nissim, "Bounds on the sample complexity for private learning and private data release," *Machine Learning*, vol. 94, no. 3, pp. 401–437, 2014.

[100] S. Vadhan, "The complexity of differential privacy," in *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, Y. Lindell, Ed. Cham, Switzerland: Springer International Publishing AG, 2017, ch. 7, pp. 347–450.

[101] T. Steinke and J. Ullman, "Interactive fingerprinting codes and the hardness of preventing false discovery," in *Proceedings of the 28th Annual Conference on Learning Theory*, ser. COLT '15, 2015, pp. 1588–1628.

[102] G. Valiant and P. Valiant, "A CLT and tight lower bounds for estimating entropy," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 17, no. 179, 2010.

[103] M. Anthony, "Classification by polynomial surfaces," *Discrete Applied Mathematics*, vol. 61, no. 2, pp. 91–103, 1995.

[104] L. Devroye, A. Mehrabian, and T. Reddad, "The total variation distance between high-dimensional Gaussians," *arXiv preprint arXiv:1810.08693*, 2018.

[105] C. Daskalakis and C. H. Papadimitriou, "On oblivious PTAS's for Nash equilibrium," in *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, ser. STOC '09. New York, NY, USA: ACM, 2009, pp. 75–84.

[106] ——, "Sparse covers for sums of indicators," *Probability Theory and Related Fields*, vol. 162, no. 3, pp. 679–705, 2015.

[107] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R. A. Servedio, and L. Y. Tan, "Learning sums of independent integer random variables," in *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 217–226.

[108] C. Daskalakis and C. H. Papadimitriou, "Discretized multinomial distributions and Nash equilibria in anonymous games," in *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, ser. FOCS '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 25–34.

[109] ——, "Approximate Nash equilibria in anonymous games," *Journal of Economic Theory*, vol. 156, pp. 207–245, 2015.

[110] R. J. Duffin and A. C. Schaeffer, "A refinement of an inequality of the brothers Markoff," *Transactions of the American Mathematical Society*, vol. 50, no. 3, pp. 517–528, 1941.

[111] D. Coppersmith and T. J. Rivlin, "The growth of polynomials bounded at equally spaced points," *SIAM Journal on Mathematical Analysis*, vol. 23, no. 4, pp. 970–983, 1992.

[112] N. Alon, R. Livni, M. Malliaris, and S. Moran, "Private PAC learning implies finite Littlestone dimension," in *Proceedings of the 51st Annual ACM Symposium on the Theory of Computing*, ser. STOC '19. New York, NY, USA: ACM, 2019, pp. 852–860.

[113] A. Beimel, K. Nissim, and U. Stemmer, "Private learning and sanitization: Pure vs. approximate differential privacy," *Theory of Computing*, vol. 12, no. 1, pp. 1–61, 2016.

Mark Bun is an Assistant Professor in the Department of Computer Science at Boston University. He completed his Ph.D. in the Theory of Computation group at Harvard University, where he was advised by Salil Vadhan. After that, he was a postdoctoral researcher at Princeton University and a Google Research Fellow at the Simons Institute for the Theory of Computing. His research interests include computational complexity, differential privacy, cryptography, and learning theory.

Gautam Kamath is an assistant professor at the David R. Cheriton School of Computer Science at the University of Waterloo. He has a B.S. in Computer Science and Electrical and Computer Engineering from Cornell University, and an M.S. and Ph.D. in Computer Science from the Massachusetts Institute of Technology. His research interests lie in principled methods for statistics and machine learning, with a focus on settings which are common in modern data analysis, including privacy and robustness. He was a Microsoft Research Fellow, as a part of the Simons-Berkeley Research Fellowship Program at the Simons Institute for the Theory of Computing. He was awarded an NSERC Discovery Accelerator Supplement, and the Best Student Presentation Award at the ACM Symposium on Theory of Computing.

Thomas Steinke was a Research Staff Member at the IBM Almaden Research Center during the course of this work. He is now a Research Scientist at Google Research in the Brain Privacy and Security Team in Mountain View, California, USA. During the early part of this work, he was also visiting the Simons Institute for the Theory of Computing at UC Berkeley as a Patrick J. McGovern Research Fellow in the Data Privacy: Foundations and Applications program. His research focuses on privacy – specifically, differential privacy – and its connections to other areas, particularly machine learning and generalization. In 2016, he completed his PhD in Computer Science at Harvard University advised by Salil Vadhan. Before that he completed his undergraduate studies and a masters degree in Mathematics at the University of Canterbury in Christchurch, New Zealand.

Zhiwei Steven Wu was an Asistant Professor in the Computer Science and Engineering Department at the University of Minnesota during the course of this work. He is now an Assistant Professor in the School of Computer Science at Carnegie Mellon University. During the early part of this work, he was a visiting scientist at the Simons Institute for the Theory of Computing at UC Berkeley in the "Data Privacy: Foundations and Applications" program. His research focuses on (1) how to make machine learning better aligned with societal values, especially privacy and fairness, and (2) how to make machine learning more reliable and robust when algorithms interact social and economic dynamics. Previously, he completed his Ph.D. in computer science at the University of Pennsylvania in 2017. His research has been generously supported by the National Science Foundation (NSF), an Amazon Research Award, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, and a Mozilla Research Grant.