Learning Poisson Binomial Distributions with Differential Privacy

National and Kapodistrian University of Athens
Department of Mathematics

Giannakopoulos Agamemnon

supervised by:
Dimitris Fotakis

March 1, 2017

Η παρούσα Διπλωματική Εργασία

εκπονήθηκε στα πλαίσια των σπουδών

για την απόκτηση του

Μεταπτυχιακού Διπλώματος Ειδίκευσης

στη

Λογική και Θεωρία Αλγορίθμων και Υπολογισμού

που απονέμει το

Τμήμα Μαθηματικών

του

Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών

Εγκρίθηκε την …………… από Εξεταστική Επιτροπή

αποτελούμενη από τους:

<u>**Ονοματεπώνυμο**</u>	<u>**Βαθμίδα**</u>	<u>**Υπογραφή**</u>
1. ………………………...	…………………………………………………	……………………
2. ………………………..	…………………………………………………	……………………
3. ………………………...	………………………………………………….	……………………

To my wife Sylvia,
and my daughter Iphigenia

Abstract

Daskalakis and Papadimitriou [7] show that for all $n, \epsilon > 0$ the set of Poisson Binomial Distributions admits a proper ϵ-cover in total variation distance, of size $n^2 + n \cdot (1/\epsilon)^{O(log^2(1/\epsilon))}$, which can also be computed in polynomial time. More specific, they proved that given a set S of PBDs there exists a ϵ-cover S_ϵ with the following property: If a PBD $X \in S$ then there exist a distribution $Y \in S_\epsilon$ that is ϵ-close to X and Y is on k-Sparse form or on (n, k)-Binomial form. Based on that theorem Daskalakis, Diakonikolas and Servedio [6] gave a highly efficient algorithm which learns to ϵ-accuracy (with respect to the total variation distance) using $O(1/\epsilon^3)$ samples independent of n. The running time of the algorithm is quasilinear in the size of its input data $O(log(n)/\epsilon^3)$. Their second main result is a proper learning algorithm that learns to ϵ-accuracy using $O(1/\epsilon^2)$ samples, and runs in time $(1/\epsilon)^{poly(log(1/\epsilon))} \cdot \log n$. The algorithm output its result in 3 stages. On stage 1 the algorithm outputs a hypothesis distribution H_S that is ϵ-close to the initial PBD X when the latter is close to a k-Sparse form in the cover S_ϵ. On stage 2 the algorithm outputs a hypothesis distribution H_P that is ϵ-close to the initial PBD X when the latter is close to a (n, k)-Binomial form in the cover S_ϵ. Finally the algorithm choose the closer distribution to X between H_S and H_P.

We try to study the above mentioned algorithm regarding the property of differential privacy. More specific we prove that if the PBD X is ϵ-close to a (n, k)-Binomial form then the algorithm is differential private. On case where X is ϵ-close to a k-Sparse form the property of differential privacy

depends on the PBD cardinality.

Acknowledgements

In my sort journey in this postgraduate programme i have met lot of great people that i would like to thank. First my fellow students, with whom we spend a lot of time studying together and helped me to be developed as a student but also as a person. All of my teachers who despite the harsh conditions, they found the strength and the will to teach and educate us. Finally, i would like to thank my supervisor Dimitrio Fotaki who he believed in me and had the patience to endure me as his student.

Contents

Chapter 1

Introduction

Due to extensive use of computers, information can be stored massively in databases. Corporations, organizations and governments collect digital information and provide a unique opportunity for the conduction of fruitful statistical research. However, their analysis may pose risks and difficulties. A major problem that usually researchers meet is their raw and unstructured form, making the extraction of useful information very challenging. The second problem is generated by the fact that most of the collected datasets contain private or sensitive information and their use may pose risks for privacy violations. Regarding the first, the area of Distribution Learning Theory has given significant results while for the second Differential Privacy tackles successfully the issue. These two areas constitute the main aspects of this thesis. Let's first introduce to the reader these two fields.

1.1 Distribution Learning Theory

As stated above a crucial problem that may arises, regarding the use of a database, is their unstructured form, which makes difficult the extraction of useful information. More specific, assuming that a large class of these datasets can be modeled as samples from a probability distribution over a very large domain, an important goal, regarding the exploration of these datasets, would be the understanding of their underlying distributions. The field which tackle this problem is known as distribution learning.

Distribution learning theory is a framework of machine learning theory and the goal is to find an efficient algorithm that, based on a known sample, determines with high probability the distribution from which the samples have been drawn. It is a recently developed area, and blends with parallel developments in computer science, and in particular machine learning. Especially after the explosion of “Big Data” problems, the specific area has become very active and with many applications in fields such as Medical Diagnosis, Finance, neural networks and many others.

Distribution learning problems have often been investigated in the context of supervised or unsupervised learning. By the term supervised learning we refer to a machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset is properly labeled by a “supervisor” or “teacher” and includes input data and response values (output data). From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset. By the term unsupervised learning we usually refer to problems were, given a training sample (input data) we try to model the underlying data structure/distribution with zero-knowledge for the output (i.e. no knowledge regarding the underlying distribution). In this thesis we consider a basic problem in unsupervised learning: learning an unknown Poisson Binomial Distribution (PBD).

PBDs are one of the most basic classes of discrete distributions. Indeed, they are arguably the simplest n parameter probability distribution that has some nontrivial structure. As such they have been intensely studied in probability and statistics and arise in many settings; for example, note here that tail bounds on PBDs form an important special case of Chernoff/Hoeffding bounds [37, 38, 39]. In application domains, PBDs have many uses in research areas such as survey sampling, case-control studies, and survival analysis, see e.g., [40] for a survey of the many uses of these distributions in applications.

Before further investigate the area of Poisson Binomial Distributions let us introduce an example to better understand the concept and usage of the distribution learning theory. Consider, for example the identification of high risk groups in a population. A physician may wish, on the basis of records of patients effected by some disease, to infer the attribute values of the subgroups of the population which are at high risk of contracting this disease. One may assume that the overall distribution of the population over the attributes space is known to the researcher and serves as a baseline relative to which risk (i.e. the density of the distribution of sick people) is defined. Note that in the situation we consider here the physician has access to files of sick people only. Consequently we may view his data as a sample drawn from the unknown distribution that he wishes to assess. Distribution learning theory could extract, with high accuracy, the unknown distribution, deriving the attribute values of the population which are at high risk of contracting the disease.

1.2 Differential Privacy

Consider the above mentioned example. To perform the statistical analysis (and define the population which are at high risk of contracting this disease) the physician is getting access to sensitive information about a group of patients. This access may be pose risks by violating patient's privacy and exposing (unwittingly) their personal information in public. Even if the physician apply some simple anonymization techniques, such as deleting user name or ID number to preserve privacy, individuals sensitive information still having a high probability of being re-identified from the released dataset. In the early years, Latanya Sweeney et al. provided an example of de-anonymization on a published medical dataset [12,13].

Lots of literatures suggests that with the background information (i.e. zip code, date of birth, gender etc.) the combination of several attributes may re-identify an individual. Narayanan [15] re-identified part of the users in the Netflix Prize dataset by associating it with the International Movie Data Base (IMDB). Mudhakar et al. [16] de-anonymized mobility traces by using social networks as their background information. Stevens et al. [17] exploited Skype, a popular P2P communication software to invade users' location and sharing of information. All of these examples show that simple anonymization is insufficient for privacy preserving. All these incidents motivated this area for more robust and efficient results. But what is the true meaning of privacy?

Privacy or better "Differential privacy" describes a promise, made by a data holder, or curator, to a data subject: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available." At their best, differentially private database mechanisms can make confidential data widely available for accurate data analysis, without resorting to data clean rooms, data usage agreements, data protection plans, or restricted views. Thus differential privacy addresses the paradox of learning nothing about an individual while learning useful information about a population. A rich class of application areas exists regarding differential privacy. Medical databases, Homeland Security, Transportation Systems, Network searching and Searching Engine companies are some indicative areas were differential privacy is applied. More specific, a medical database may give us a lot of information regarding a specific disease (i.e. smoking causes cancer), without exposing sensitive information of the individuals who constitute the dataset.

Differential privacy acquires the intuition that releasing an aggregated report should not reveal too much information on any individual record in the dataset [14]. This can be achieved using randomized mechanisms whose output distribution remains almost unchanged even with an arbitrary individual record deleted. More precisely, the randomized mechanism contains adding calibrated noise to the query output or randomizing all possible outputs in the domain.

Differential privacy is a definition, not an algorithm. For a given computational task T and a given value of ϵ there will be many differentially private algorithms for achieving T in an ϵ-differentially private manner. Some will have better accuracy than others. When ϵ is small, finding a highly accurate ϵ-differentially private algorithm for T can be difficult, much as finding a numerically stable algorithm for a specific computational task can require effort.

Distribution Learning and Differential Privacy are the key aspects that will be examined in this thesis. Having get a glimpse of them lets continue with some historical points and the importance of their applications.

1.3 Historical Background and Applications

Regarding Distribution Learning theory the first pioneer was Karl Pearson, 1894 [1] who tried to estimate the parameters of a mixture of Gaussians (a linear combination of two Gaussian distributions). Pearson was the first who introduced the notion of the mixtures of Gaussians in his attempt to explain the probability distribution from which he got same data that he wanted to analyze. The learning procedure starts with clustering the samples into two different clusters minimizing some metric. Using the assumption that the means of the Gaussians are far away from each other with high probability the samples in the first cluster correspond to samples from the first Gaussian and the samples in the second cluster to samples from the second one. Applications for this work are met in areas such as Fisheries research, Agriculture, Botany, Economics, Medicine, Genetics, Psychology, Paleontology, Electrophoresis, Finance, Sedimentology/Geology and Zoology [2]. A range of new methods regarding mixture Gaussian estimation have been proposed. More specific Valiant Moitra and Kalai [29] proposed a polynomial-time algorithm for the case of two Gaussians in n dimensions (even if they overlap), with provably minimal assumptions on the Gaussians, and polynomial data requirements. In statistical terms, the estimator converges at an inverse polynomial rate.

Valiant has also contribute in the development of the specific area by his publication [30] which shows that it is possible to design learning machines that have all three of the following properties: 1. The machines can provably learn whole classes of concepts. Furthermore, these classes can be characterized. 2. The classes of concepts are appropriate and nontrivial for general-purpose knowledge. 3. The computational process by which the machines deduce the desired programs requires a feasible (i.e., polynomial) number of steps.

A significant work on the specific area has also be made by Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert Schapire and Linda Sellie in 1994 [3]. In their model, an unknown target distribution is chosen from a known restricted class of distributions over $\{0,1\}^n$, and the learning algorithm receives independent random draws from the target distribution. The algorithm also receives a confidence parameter δ and an approximation parameter ϵ. The goal is to output with probability at least $1-\delta$, and in polynomial time, a hypothesis distribution which has distance at most ϵ to the target distribution. Lucien Birge 1997 [4] has given a computationally efficient algorithm that can learn any unimodal distribution over a domain $[a,b]$ to variation distance ϵ from $O(log(n)/\epsilon^3)$ samples. Chan, Diakonikolas, Servedio, and X. Sun have also contribute in the development of the area by their papers [31,32,33]. Briefly, in these papers they present efficient algorithms that approximate specific classes of distributions (i.e. a class C of probability distributions over the discrete domain $[n]$ that can be well-approximated by a variable-width histogram with few bins; univariate probability distributions that are well approximated by piecewise polynomial density functions; distributions with a piecewise constant probability density function).

Finally extensive work has also been made by Papadimitriou, Daskalakis, Diakonikolas and Servedio in their respective papers [5,6,7,8]. Summarizing their results, they proposed a highly efficient algorithm which learns to ϵ-accuracy (with respect to the total variation distance) using $O(1/\epsilon^3)$ samples independent of n. The running time of the algorithm is quasilinear in the size of its input data, i.e., $O(log(n)/\epsilon^3)$ bit-operations. Their second main result is a proper learning algorithm that learns to ϵ-accuracy using $O(1/\epsilon^2)$ samples, and runs in time $(1/\epsilon)^{poly(log(1/\epsilon))} \cdot logn$. This sample complexity is nearly optimal, since any algorithm for this problem must use $\Omega(1/\epsilon^2)$ samples. The results of the specific papers will be further analyzed in the next chapters. Distribution learning as a framework of machine learning theory may appear in a huge area of applications as Medicine, Neural Networks, Finance and many others areas [9,10,11].

Regarding Differential Privacy. Various methods and algorithms have been proposed to preserve privacy. The most popular privacy model is the k-anonymity model 1996 [20, 21]. It partitions the dataset into a number of equivalence groups in which every record has the same attribute values with at least $K - 1$ other records. There are a number of other privacy models. For example, the ℓ-diversity 2007 [22] privacy model ensures at least l diverse values exist for the sensitive attribute. The t-closeness 2007 [23,24] model requires the sensitive attribute distribution in each group should not deviate from that of the whole dataset by more than t. The δ-presence 2007 [25] bounds the probability of inferring the presence of any individual's record within a specified range.

The following algorithms are among the most known and classical differential private algorithms. Dwork, McSherry, Nissim, and Smith 2006 first introduce The Laplace mechanism [18]. On input a query function f mapping databases to reals, the so-called true answer is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution (Laplace distribution), and this response, the true answer plus noise, is returned to the user. McSherry and Talwar 2009 [19] proposed the exponential mechanism. Given some arbitrary range R, the exponential mechanism is defined with respect to some utility function u, which maps database/output pairs to utility scores. Intuitively, for a fixed database x, the user prefers that the mechanism outputs some element of R with the maximum possible utility score.

Applications of differential private mechanisms/algorithms may be found in each area were databases with sensitive information exists. Some indicative examples are given bellow:

Medical Databases: The scrub system [26] was designed for identification of clinical notes and letters which typically occurs in the form of textual data. The Scrub system uses numerous detection algorithms which compete in parallel to determine when a block of text corresponds to a name, address or a phone number. The Datafly System [27] was one of the earliest practical applications of privacy-preserving transformations. This system was designed to prevent identification of the subjects of medical records which may be stored in multidimensional format.

Homeland Security Applications: A number of applications for homeland security are inherently intrusive because of the very nature of surveillance. In [28], a broad overview is provided on how privacy-preserving techniques may be used in order to deploy these applications effectively without violating user privacy.

Montreal Transportation System: With the wide deployment of smart card automated fare collection (SCAFC) systems, public transit agencies have been benefiting from huge volume of transit data, a kind of sequential data, collected every day. Yet, improper publishing and use of transit data could jeopardize passengers' privacy. R. Chen, B. C. Fung, B. C. Desai, and

N. M. Sossou present in their paper [34], a solution to transit data publication under the rigorous differential privacy model for the Soci de transport de Montr (STM). They propose an efficient data-dependent yet differentially private transit data sanitization approach based on a hybrid-granularity prefix tree structure.

Search engine companies: Search engine companies collect the database of intentions, the histories of their users' search queries. These search logs are a gold mine for researchers. Search engine companies, however, are wary of publishing search logs in order not to disclose sensitive information. M. Gotz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke in their paper [35], are analyzing algorithms for publishing frequent keywords, queries, and clicks of a search log. They first show how methods that achieve variants of k-anonymity are vulnerable to active attacks and demonstrate that the stronger guarantee ensured by ϵ-differential privacy unfortunately does not provide any utility for this problem. However they propose an algorithm (ZEALOUS) and show how to set its parameters to achieve (ϵ, δ)-probabilistic privacy.

Netflix Prize: F. McSherry and I. Mironov in their paper [36] are consider the problem of producing recommendations from collective user behavior while simultaneously providing guarantees of privacy for these users. Specifically, they consider the Netflix Prize data set, and its leading algorithms, adapted to the framework of differential privacy. Differential privacy constrains a computation in a way that precludes any inference about the underlying records from its output. Such algorithms necessarily introduce uncertainty (i.e. noise) to computations, trading accuracy for privacy. They find that several of the leading approaches in the Netflix Prize competition can be adapted to provide differential privacy, without significantly degrading their accuracy. To adapt these algorithms, they explicitly factor them into two parts, an aggregation/learning phase that can be performed with differential privacy guarantees, and an individual recommendation phase that uses the learned correlations and an individual's data to provide personalized recommendations.

1.4 Motivation / Results

In this thesis we consider a basic problem in unsupervised learning: learning an unknown Poisson Binomial Distribution (PBD). Daskalakis, Diakonikolas and Servedio [6] proposed an algorithm for efficient for learning (PBD) from a known sample. We try to prove differential privacy on [6] algorithm, regarding Poisson Binomial Distribution learning. More specific we prove that if the PBD X is ϵ-close to a (n, k)-Binomial form then the algorithm is differential private. On case where X is ϵ-close to a k-Sparse form the property of differential privacy depends on the PBD cardinality.

Chapter 2

Preliminaries

As mentioned above Distribution Learning and Differential Privacy are the key aspects that will be examined in this thesis. Thus, a detailed analysis on the prerequisites definitions/theorems for both areas are given in the next subsections. We further need the following notation.

Order Notation: Whenever we write $O(f(n))$ or $\Omega(f(n))$ in some bound where n ranges over the integers, we mean that there exists a constant $c > 0$ such that the bound holds true for sufficiently large n if we replace the $O(f(n))$ or $\Omega(f(n))$ in the bound by $c \cdot f(n)$. On the other hand, whenever we write $O(f(1/\epsilon))$ or $\Omega(f(1/\epsilon))$ in some bound where ϵ ranges over the positive reals, we mean that there exists a constant $c > 0$ such that the bound holds true for sufficiently small ϵ if we replace the $O(f(1/\epsilon))$ or $\Omega(f(1/\epsilon))$ in the bound with $c \cdot f(1/\epsilon)$.

2.1 Poisson Binomial Distribution

Total Variation Distance: For two distributions P and Q supported on a finite set A their total variation distance is defined as:

$$d_{TV}(P,Q) := (1/2) \cdot \sum_{\alpha \in A} |P(a) - Q(a)|.$$

Covers: Let F be a set of probability distributions. A subset $G \subseteq F$ is called a (proper) ϵ-cover of F in total variation distance if, for all $D \in F$, there exists some $D' \in G$ such that $d_{TV}(D, D') \leq \epsilon$.

Poisson Binomial Distribution: A Poisson Binomial distribution of order $n \in \mathbb{N}$ is the discrete probability distribution of the sum $\sum_{i=1}^{n} X_i$ of n mutually independent Bernoulli random variables $X_1, ..., X_n$.

We denote the set of all Poisson Binomial distributions of order n by S_n. By definition, a Poisson Binomial distribution $D \in S_n$ can be represented by a vector $(p_i)_{i=1}^{n} \in [0,1]^n$ of probabilities as follows. We map $D \in S_n$ to a vector of probabilities by finding a collection $X_1, ..., X_n$ of mutually independent indicators such that $\sum_{i=1}^{n} X_i$ is distributed according to D, and setting $p_i = \mathbb{E}[X_i]$ for all i. We will be denoting a Poisson Binomial distribution $D \in S_n$ by $PBD(p_i, ..., p_n)$ when it is on the latter form. Lemma 2.1 implies that the resulting vector of probabilities is unique up to a permutation, so that there is a one-to-one correspondence between Poisson Binomial distributions and vectors $(p_i)_{i=1}^{n} \in [0,1]^n$ such that $0 \leq p_i \leq p_2 \leq ... \leq p_n \leq 1$.

Translated Poisson Distribution: We say that an integer random variable Y is distributed according to the translated Poisson distribution with parameters μ and σ^2, denoted $TP(\mu, \sigma^2)$, iff Y can be written as

$$Y = \lfloor \mu - \sigma^2 \rfloor + Z,$$

where Z is a random variable distributed according to $Poisson(\sigma^2 + \{\mu - \sigma^2\})$, where $\{\mu - \sigma^2\}$ represents the fractional part of $\mu - \sigma^2$.

Lemma 2.1. *Let $X_1, ..., X_n$ be mutually independent indicators with expectations $p_1 \leq p_2 \leq ... \leq p_n$ respectively. Similarly let $Y_1, ..., Y_n$ be mutually independent indicators with expectations $q_1 \leq q_2 \leq ... \leq q_n$ respectively. The distributions of $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n Y_i$ are different if and only if $(p_1, ..., p_n) \neq (q_1, ..., q_n)$.*

Lemma 2.2. *Let $X_1, ..., X_n$ be mutually independent random variables, and let $Y_1, ..., Y_n$ be mutually independent random variables. Then*

$$d_{TV}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\right) \leq \sum_{i=1}^n d_{TV}(X_i, Y_i).$$

Lemma 2.3. *(Variation distance of Poisson Distributions). Let $\ell_1, \ell_2 > 0$. Then*

$$d_{TV}(Poisson(\ell_1), Poisson(\ell_2)) \leq \frac{1}{2}(e^{|\ell_1 - \ell_2|} - e^{-|\ell_1 - \ell_2|}).$$

Lemma 2.4. *(Variation distance of Translated Poisson Distributions [49]) Let $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in \mathbb{R}_+$ be such that $\lfloor \mu_1 - \sigma_1^2 \rfloor \leq \lfloor \mu_2 - \sigma_2^2 \rfloor$. Then*

$$d_{TV}(TP(\mu_1, \sigma_1^2), TP(\mu_2, \sigma_2^2)) \leq |\mu_1 - \mu_2|\sigma_1 + \frac{|\sigma_1^2 - \sigma_2^2| + 1}{\sigma_1^2}.$$

Theorem 2.5. *(Binomial Approximation [50]). Let $J_1, ..., J_n$ be mutually independent indicators with $\mathbb{E}[J_i] = t_i$, and $\bar{t} = \frac{\sum_i t_i}{n}$. Then*

$$d_{TV}\left(\sum_{i=1}^n J_i, \mathcal{B}(n, \bar{t})\right) \leq \frac{\sum_{i=1}^n (t_i - \bar{t})^2}{(n+1)\bar{t}(1-\bar{t})},$$

where $\mathcal{B}(n, \bar{t})$ is the Binomial distribution with parameters n and $\bar{t}$

Theorem 2.6. *(Translated Poisson Approximation[43]). Let $J_1, ..., J_n$ be mutually independent indicators with $\mathbb{E}[J_i] = t_i$. Then*

$$d_{TV}\left(\sum_{i=1}^n J_i, TP(\mu, \sigma^2)\right) \leq \frac{\sqrt{\sum_{i=1}^n t_i^3(1-t_i)} + 2}{\sum_{i=1}^n t_i(1-t_i)},$$

where $\mu = \sum_{i=1}^n t_i$ and $\sigma^2 = \sum_{i=1}^n t_i(1-t_i)$.

Theorem 2.7. *(Poisson Approximation). Let $J_1, ..., J_n$ be mutually independent indicators with $\mathbb{E}[J_i] = t_i$, and $\bar{t} = \frac{\sum_i t_i}{n}$. Then*

$$d_{TV}\left(\sum_{i=1}^{n} J_i, Poisson\left(\sum_i t_i\right)\right) \leq \frac{\sum_i t_i^2}{\sum_i t_i}.$$

Theorem 2.8. *(Chernoff Bounds). Let $X = \sum_{i=1}^{n} X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}(X) = \sum_{i=1}^{n} p_i$. Then*

$$\mathbb{P}(|X - \mu| \geq \delta\mu) \leq 2e^{\mu\delta^2/3} \quad \textit{for all} \quad 0 < \delta < 1.$$

2.2 Differential Privacy

Distance Between Databases: The ℓ_1 norm of a database x is denoted $\|x\|_1$ and is defined to be:

$$\|x\|_1 = \sum_{i=1}^{|X|} |x_i|$$

The ℓ_1 distance between two databases x and y is $\|x - y\|_1$. Note that $\|x\|_1$ is a measure of the size of a database x (i.e., the number of records it contains), and $\|x - y\|_1$ is a measure of how many records differ between x and y. Databases may also be represented by multisets of rows (elements of X) or even ordered lists of rows, which is a special case of a set, where the row number becomes part of the name of the element. In this case distance between databases is typically measured by the Hamming distance, i.e., the number of rows on which they differ.

Differential Privacy: A randomized algorithm M with domain $N^{|X|}$ is ϵ-differential private if for all $S \subseteq Range(M)$ and for all $x, y \in N^{|X|}$ such that $\|x - y\|_1 \leq 1$:

$$Pr[M(x) \in S] \leq \exp(\epsilon) \cdot Pr[M(y) \in S],$$

where, $\|x\|_1$ the ℓ_1 norm of a database x. Substituting the term $\exp(\epsilon)$ by $1-\epsilon$ the above relation may be expressed as

$$Pr[M(x) \in S]/Pr[M(y) \in S] \leq 1+\epsilon.$$

Intuitively the above relation tell us that if an algorithm is differential private, then a small change on his input database (by at most one entry) will affect negligible its output. Thus if an individual adds his personal information on a predefined database the algorithm's output will not change, betraying his information.

To better understand the concept of differential privacy we will illustrate a classical mechanism (the Laplace Mechanism) regarding the latter.

Definition 2.1. *(ℓ_1 sensitivity). The ℓ_1 sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$ is:*

$$\Delta f = \underset{\substack{\|x-y\|_1=1 \\ x,y \in \mathbb{N}^{|\mathcal{X}|}}}{max} \parallel f(x) - f(y) \parallel_1 .$$

The ℓ_1 sensitivity of a function f captures the magnitude by which a single individual's data can change the function f in the worst case, and therefore, intuitively, the uncertainty in the response that we must introduce in order to hide the participation of a single individual. Indeed, we will formalize this intuition: the sensitivity of a function gives an upper bound on how much we must perturb its output to preserve privacy. One noise distribution naturally lends itself to differential privacy.

Definition 2.2. *(The Laplace Distribution). The Laplace Distribution (centered at 0) with scale b is the distribution with probability density function*

$$Lap(x|b) = \frac{1}{2b} exp(-\frac{|x|}{b}).$$

The variance of this distribution is $\sigma^2 = 2b^2$. We will sometimes write $Lap(b)$ to denote the Laplace distribution with scale b, and will sometimes abuse notation and write *Lap*(b) simply to denote a random variable $X \sim$ *Lap*(b).

We will now define the Laplace Mechanism. As its name suggests, the Laplace mechanism will simply compute f, and perturb each coordinate with noise drawn from the Laplace distribution. The scale of the noise will be calibrated to the sensitivity of f (divided by ϵ).

Definition 2.3. *(Laplace Mechanism). Give any function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$, the Laplace mechanism is defined as:*

$$\mathcal{M}_L(x, f(\cdot), \epsilon) = f(x) + (Y_1, ..., Y_k)$$

where Y_i are i.i.d. random variables drawn from $Lap(\Delta f/\epsilon)$.

Theorem 2.9. *The Laplace mechanism is ϵ-differential private.*

Proof. Let $x \in \mathbb{N}^{|\mathcal{X}|}$ and $y \in \mathbb{N}^{|\mathcal{X}|}$ be such that $\parallel x - y \parallel_1 \leq 1$, and let $f(\cdot)$ be some function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^k$. Let p_x denote the probability density function of $\mathcal{M}_L(x, f, \epsilon)$, and let p_y denote the probability density function of $\mathcal{M}_L(y, f, \epsilon)$. We compare the two at some arbitrary point $z \in \mathbb{R}^k$.

$$\frac{p_x(z)}{p_y(z)} = \prod_{i=1}^{k} \left(\frac{exp(-\frac{\epsilon|f(x)_i - z_i|}{\Delta f})}{exp(-\frac{\epsilon|f(y)_i - z_i|}{\Delta f})} \right)$$

$$= \prod_{i=1}^{k} exp\left(\frac{\epsilon(|f(y)_i - z_i| - |f(x)_i - z_i|)}{\Delta f} \right)$$

$$\leq \prod_{i=1}^{k} exp\left(\frac{\epsilon|f(x)_i - f(y)_i|}{\Delta f} \right)$$

$$= exp\left(\frac{\epsilon \parallel f(x) - f(y) \parallel_1}{\Delta f} \right)$$

$$\leq \exp(\epsilon),$$

where the first inequality follows from the triangle inequality, and the last follows from the definition of sensitivity and the fact that $\parallel x - y \parallel_1 \leq 1$. That $\frac{p_x(z)}{p_y(z)} \geq exp(-\epsilon)$ follows by symmetry.

Chapter 3

Sparse Covers for Sums of Indicators

In this chapter we provide two crucial theorems (from [7]) that will help us to further understand the space of PBDs. More specific the first theorem proves that for each set S_n of PBDs and each $\epsilon > 0$, there exist an ϵ-cover $S_{n,\epsilon}$ of size $n^2 + (\frac{1}{\epsilon})^{O(1/\epsilon^2)}$. The specific theorem not only guarantees the existence of such a cover, but also specifies the form of the PBDs in the cover. More specific we will see that each PBD in the cover is in k-Sparse form or in (n, k)-Binomial Form. Thus, the expected values of each indicator $\mathbb{E}[X_i]$ (recall the definition of PBDs as sums of independent indicators $\sum_i X_i$) takes specific values. The second theorem sparsifies the above mentioned cover $S_{n,\epsilon}$ by removing specific elements and without losing much of accuracy (i.e. losing ϵ).

Theorem 3.1. [7]. *Let $X_1, ..., X_n$ be an arbitrary mutually independent indicators, and $k \in \mathbb{N}$. Then there exist mutually independent indicators $Y_1, ..., Y_n$ satisfying the following:*

1. $d_{TV}\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i\right) \leq 41/k$.

2. *at least one of the following is true:*

(a) *($k - sparseform$) there exists some* $\ell \leq k^3$ *such that, for all* $i \leq \ell, \mathbb{E}[Y_i] \in \left\{\frac{1}{k^2}, \frac{2}{k^2}, ..., \frac{k^2-1}{k^2}\right\}$ *and, for all* $i < l, \mathbb{E}[Y_i] \in \{0,1\}$*;or*

(b) *($(n,k) - Binomialform$) there is some* $\ell \in \{1, ..., n\}$ *and* $q \in \{1, ..., n\}$ *such that, for all* $i < \ell, \mathbb{E}[Y_i] = q$ *and for all* $i > \ell, \mathbb{E}[Y_i] = 0$*; moreover,* ℓ *and* q *satisfy* $\ell q \geq k^2$ *and* $\ell q(1-q) \geq k^2 - k - 1$.

Proof. Assume that $\mathbb{E}[X_i] = p_i$ thus the vector $(p_1, ..., p_n)$ corresponds the expectation values of X_i's. The proof of Theorem is conducted in 2 stages however, the high level of the proof is the following: We will allocate the values of all $p_i \in (0,1)$ to specific values in such a way so we do not have to travel too much distance from the starting Poisson Binomial distribution. The details of the proof are presented hereafter.

In Stage 1 we allocate the probabilities p_i's $\in (0, \frac{1}{k}) \cup (1 - \frac{1}{k}, 1)$ to the discrete values $0, \frac{1}{k}, 1 - \frac{1}{k}$ and 1. This allocation will yield a new vector of probabilities $p_i^{'}$ and thus the creation of new variables $Z_1, ..., Z_n$ with $\mathbb{E}[Z_i] = p_i^{'}$ such that

$$d_{TV}\left(\sum_i X_i, \sum_i Z_i\right) \leq 7/k$$

and

$$\mathbb{E}[Z_i] \notin (0, \frac{1}{k}) \cup (1 - \frac{1}{k}, 1)$$

that is we eliminate from our collection variables that have expectations very close to 0 and 1, without traveling to much from the initial PBD. To achieve this bound the following steps are applied:

We first define $\mathcal{L}_k$ be the set of all i's such that $p_i \in (0, 1/k)$ and $\mathcal{H}_k$ be the set of all i's such that $p_i \in (1 - \frac{1}{k}, 1)$.

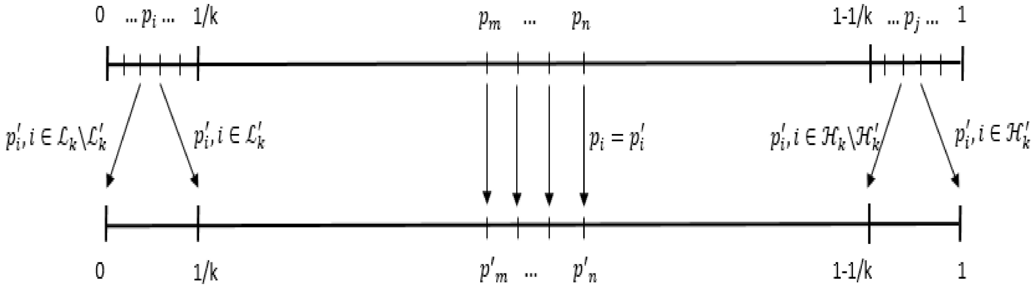
1. Initially, we set $p_i^{'} = p_i$, for all $i \in [n] \backslash \mathcal{L}_k \cup \mathcal{H}_k$. We simple set $\mathbb{E}[Z_i] = p_i$ for all $p_i \in [\frac{1}{k}, 1 - \frac{1}{k}]$ thus,

$$d_{TV}\left(\sum_{i \in [n] \backslash \mathcal{L}_k \cup \mathcal{H}_k} X_i, \sum_{i \in [n] \backslash \mathcal{L}_k \cup \mathcal{H}_k} Z_i\right) = 0.$$

2. Let $\mathcal{L}_k^{'} \subseteq \mathcal{L}_k$ be an arbitrary subset of cardinality $r = \lfloor \frac{\sum_{i \in L_k} p_i}{1/k} \rfloor$. We simply identify (randomly) a set of indicators i such that to round all the p_i's in $(0, 1/k)$ to 0 and $1/k$

3. Set $p_i^{'} = \frac{1}{k}$, for all $i \in \mathcal{L}'_k$, and $p_i^{'} = 0$, for all $i \in \mathcal{L}_k \backslash \mathcal{L}'_k$.

We round all indicators expectations to 0 or $1/k$. The following figure shows the Stage 1 process.


0
$\dots p_i \dots$
1/k
p_m
$\dots$
p_n
1-1/k
$\dots p_j \dots$
1
$p'_i, i \in \mathcal{L}_k \backslash \mathcal{L}'_k$
$p'_i, i \in \mathcal{L}'_k$
$p_i = p'_i$
$p'_i, i \in \mathcal{H}_k \backslash \mathcal{H}'_k$
$p'_i, i \in \mathcal{H}'_k$
0
1/k
p'_m
$\dots$
p'_n
1-1/k
1

The first 3 steps finalize Stage 1 process. Let us define how much this rounding costs in terms of variation distance. The first step implies zero distance as we saw above. For steps 2-3 we will bound the distance $d_{TV}(\sum_{i \in \mathcal{L}_k} X_i, \sum_{i \in \mathcal{L}_k} Z_i)$ using Theorem 2.7. In particular

$$d_{TV}\left(\sum_{i \in \mathcal{L}_k} X_i, Poisson\left(\sum_{i \in \mathcal{L}_k} p_i\right)\right) \leq \frac{\sum_{i \in \mathcal{L}_k} p_i^2}{\sum_{i \in \mathcal{L}_k} p_i} \leq \frac{\frac{1}{k}\sum_{i \in \mathcal{L}_k} p_i}{\sum_{i \in \mathcal{L}_k} p_i} = 1/k.$$

Similarly, $d_{TV}\left(\sum_{i \in \mathcal{L}_k} Z_i, Poisson\left(\sum_{i \in \mathcal{L}_k} p_i^{'}\right)\right) \leq 1/k$. By Lemma 2.3 we bound the distance

$$d_{TV}\left(Poisson\left(\sum_{i \in \mathcal{L}_k} p_i\right), Poisson\left(\sum_{i \in \mathcal{L}_k} p_i^{'}\right)\right) \leq \frac{1}{2}(e^{\frac{1}{k}} - e^{-\frac{1}{k}}) \leq \frac{1.5}{k}.$$

Using the triangle inequality we get that

$$d_{TV}(\sum_{i \in \mathcal{L}_k} X_i, \sum_{i \in \mathcal{L}_k} Z_i) \leq \frac{3.5}{k}.$$

Thus the Stage 1 keeps k-close the distance between the two RBDs $\sum_i X_i$ and $\sum_i Z_i$.

Stage 2 is a more complex process. In this stage the rounding of $p_i^{'}$'s (from Stage 1) it depends from the number m of $p_i^{'} \in (0,1)$. More specific if $m \leq k^3$ then we will construct indicators $Y_1, ..., Y_n$ which satisfy the Property $2(a)$ in the Theorem, if $m > k^3$ the relative indicators $Y_1, ..., Y_n$ will satisfy the Property $2(b)$ in the Theorem. We will examine both cases (a) and (b) separately. We first define $\mathcal{M}$ be all i's such that $p_i^{'} \notin \{0,1\}$ and $m := |\mathcal{M}|$.

Case (a): $m \leq k^3$

The high level of the proof is the following: We first partitioning the interval $[1/k, 1-1/k]$ into irregularly sized subintervals, whose endpoints are integer multiples of $1/k^2$. We then round **all but one** of the p_i's falling in each subinterval to the endpoints of the subinterval so as to maintain their total expectation, and apply Ehm's [50] approximation to argue that the distribution of their sum is not affected by more than $O(1/k^2)$ in total variation distance. The proof details are given hereafter

We first split the interval $[1/k, 1-1/k]$ as stated above and define the subsets of i's that fall inside each partition. More specific we first define $\mathcal{M}_l = \{i \in \mathcal{M} | p_i^{'} \leq 1/2\}$ and $\mathcal{M}_h = \{i \in \mathcal{M} | p_i^{'} \geq 1/2\}$ $(\mathcal{M} = \mathcal{M}_l \sqcup \mathcal{M}_h)$. We then define

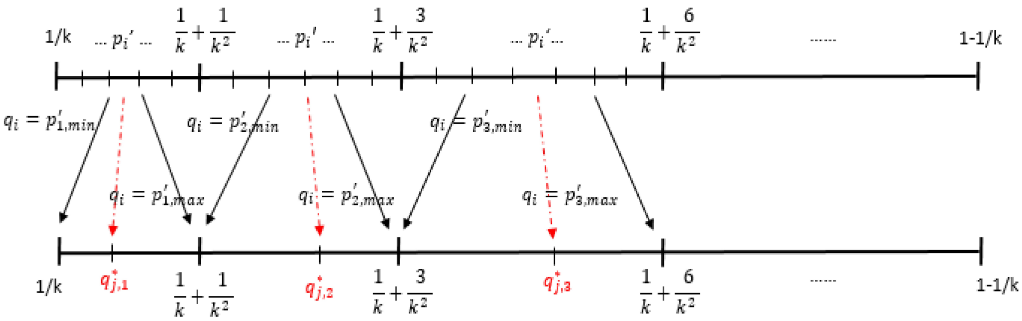
$$\mathcal{M}_{l,j} = \left\{ i | p_i^{'} \in \left[\frac{1}{k} + \frac{(j-1)j}{2} \frac{1}{k^2}, \frac{1}{k} + \frac{(j+1)j}{2} \frac{1}{k^2} \right] \right\}.$$

The same split follows regarding $\mathcal{M}_h$. We define $(q_i)_{i \in \mathcal{M}_{l,j}}$ following the next steps:

1. Set $p_{j,min} := \frac{1}{k} + \frac{(j-1)j}{2} \frac{1}{k^2}$ (the lower p_i value of the respective subinterval), $p_{j,max} := \frac{1}{k} + \frac{(j+1)j}{2} \frac{1}{k^2}$ (the maximum p_i value of the respective subinterval)

2. We define an arbitrary subset $\mathcal{M}'_{l,j} \subseteq \mathcal{M}_{l,j}$ of cardinality $r = \lfloor \frac{n_j(\bar{p}_j - p_{j,min})}{j/k^2} \rfloor$, where $n_j = |\mathcal{M}_{i,j}|$, and $\bar{p}_j = \frac{\sum_{i \in \mathcal{M}_{l,j}} p'_i}{n_j}$
3. Set $q_i = p_{j,max}$, for all $i \in \mathcal{M}'_{l,j}$,
4. for an arbitrary index $i^*_j \in \mathcal{M}_{l,j} \backslash \mathcal{M}'_{l,j}$, set $q_{i^*_j} = n_j \bar{p}_j - (r p_{j,max} + (n_j - r - 1) p_{j,min})$;
5. finally, set $q_i = p_{j,min}$, for all $i \in \mathcal{M}_{l,j} \backslash \mathcal{M}'_{l,j} \backslash \{i^*_j\}$.

The next figure shows the above mentioned procedure:



Observe that for $i \in \mathcal{M}_{l,j} \backslash i^*_j$, q_i is an integer multiple of $1/k^2$. The final step is to give an upper bound of the allocation performed in the above steps. By theorem 2.5

$$d_{TV}\left(\sum_{i \in \mathcal{M}_{l,j}} Z_i, \mathcal{B}(n_j, \bar{p}_j)\right) \leq \frac{\sum_{i \in \mathcal{M}_{l,j}} (p'_i - \bar{p}_j)^2}{(n_j + 1)\bar{p}_j(1 - \bar{p}_j)} \leq 8/k^2.$$

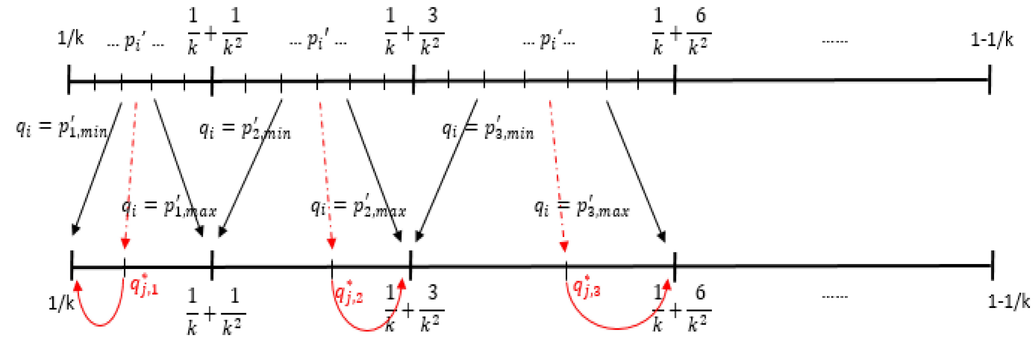
A similar deviation gives $d_{TV}\left(\sum_{i \in \mathcal{M}_{l,j}} Y_i, \mathcal{B}(n_j, \bar{p}_j)\right) \leq \frac{8}{k^2}$. Thus by triangle inequality:

$$d_{TV}\left(\sum_{i \in \mathcal{M}_{l,j}} Z_i, \sum_{i \in \mathcal{M}_{l,j}} Y_i\right) \leq \frac{16}{k^2}.$$

As the above inequality holds for each $j = 1, ..., k-1$, by lemma 2.2 we get:

$$d_{TV}\left(\sum_{i\in\mathcal{M}_l} Z_i, \sum_{i\in\mathcal{M}_l} Y_i\right) \leq \sum_{j=1}^{k-1} d_{TV}\left(\sum_{i\in\mathcal{M}_{l,j}} Z_i, \sum_{i\in\mathcal{M}_{l,j}} Y_i\right) \leq \frac{16}{k}.$$

Finally by rounding $q_{i_j^*}$ to their closest multiple of $1/k^2$ (see figure) the above bound increasing to $17/k$.



By lemma 2.2 we get the final bound of the Case (a):

$$d_{TV}\left(\sum_{i\in\mathcal{M}} Z_i, \sum_{i\in\mathcal{M}} Y_i\right) \leq d_{TV}\left(\sum_{i\in\mathcal{M}_l} Z_i, \sum_{i\in\mathcal{M}_l} Y_i\right) + d_{TV}\left(\sum_{i\in\mathcal{M}_h} Z_i, \sum_{i\in\mathcal{M}_h} Y_i\right) \leq \frac{34}{k}.$$

Case (b): $m > k^3$

The case (b) is a more technical case. At the specific approach we will approximate the PBD $\sum_i Z_i$ with a Translated Poisson distribution, using Theorem 2.6 due to Rollin [43]. The approximation of $\sum_i Z_i$, is at least $3/k$-close to the Translated Poisson distribution. We then argue that the latter is $6/k$-close to a Binomial distribution $\mathcal{B}(m', q)$, where

$$m' := \left\lceil \frac{(\sum_{i\in\mathcal{M}} p_i^{'} + t)^2}{\sum_{i\in\mathcal{M}} p_i^{'2} + t} \right\rceil \quad \text{and} \quad q := \frac{\ell^*}{n},$$

where ℓ^* satisfies $\frac{\sum_{i\in\mathcal{M}} p_i^{'}+t}{m'}$. For fixed m' and q, we set $q_i = q$, for all $i < m'$, and $q_i = 0$, for all $i > m'$. We also define:

$$\mu := \mathbb{E}\left[\sum_{i\in\mathcal{M}} Z_i\right] \quad \text{and} \quad \mu' := \mathbb{E}\left[\sum_{i\in\mathcal{M}} Y_i\right],$$

$$\sigma^2 := Var\left[\sum_{i\in\mathcal{M}} Z_i\right] \quad \text{and} \quad \sigma'^2 := Var\left[\sum_{i\in\mathcal{M}} Y_i\right].$$

The following lemma compares the values $\mu, \mu', \sigma^2, \sigma'^2$ and will provide useful information regarding proof's procedure.

Lemma 3.2. *The following hold*

$$\mu \leq \mu' \leq \mu + 1 \tag{3.1}$$

$$\sigma^2 - 1 \leq \sigma'^2 \leq \sigma^2 + 2 \tag{3.2}$$

$$\mu \geq k^2 \tag{3.3}$$

$$\sigma^2 \geq k^2(1 - \frac{1}{k}) \tag{3.4}$$

We first approximate $\sum_{i\in\mathcal{M}} Z_i$ and $\sum_{i\in\mathcal{M}} Y_i$ with a Translated Poisson distribution as follows:

$$\begin{aligned} d_{TV}\left(\sum_i Z_i, TP(\mu,\sigma^2)\right) &\leq \frac{\sqrt{\sum_i p_i^{'3}(1-p_i^{'})}+2}{\sum_i p_i^{'}(1-p_i^{'})} \leq \frac{\sqrt{\sum_i p_i^{'}(1-p_i^{'})}+2}{\sum_i p_i^{'}(1-p_i^{'})} \\ &\leq \frac{1}{\sqrt{\sum_i p_i^{'}(1-p_i^{'})}} + \frac{2}{\sum_i p_i^{'}(1-p_i^{'})} = \frac{1}{\sigma} + \frac{2}{\sigma^2} \\ &\leq \frac{1}{k\sqrt{1-1/k}} + \frac{2}{k^2(1-1/k)} \quad \text{(using (3.4))} \\ &\leq \frac{3}{k}, \end{aligned}$$

Similarly,

$$d_{TV}\left(\sum_i Y_i, TP(\mu',\sigma'^2)\right) \leq \frac{3}{k}, \quad \text{(using (3.2), (3.4))}$$

By the triangle inequality we get:

$$d_{TV}\left(\sum_i Z_i, \sum_i Y_i\right)$$

$$\le d_{TV}\left(\sum_i Z_i, TP(\mu,\sigma^2)\right) + d_{TV}\left(\sum_i Y_i, TP(\mu',\sigma'^2)\right) + d_{TV}\left(TP(\mu,\sigma^2), TP(\mu,\sigma^2)\right)$$

$$\le 6/k + d_{TV}\left(TP(\mu,\sigma^2), TP(\mu,\sigma^2)\right)$$

$$\le 6/k + \frac{|\mu-\mu'|}{min(\sigma,\sigma')} + \frac{|\sigma^2-\sigma'^2|+1}{min(\sigma^2,\sigma'^2)} \quad \text{(using lemma 2.4)}$$

$$\le \frac{1}{k\sqrt{1-\frac{1}{k}-\frac{1}{k^2}}} + \frac{3}{k^2(1-\frac{1}{k}-\frac{1}{k^2})} \quad \text{(using lemma 3.2)}$$

$$\le 9/k.$$

Theorem 3.1 implies the existence of an ϵ-cover of S_n whose size is $n^2 + n \cdot (1/\epsilon)^{O(1/\epsilon^2)}$. This cover can be obtained by enumerating over all Poisson Binomial distributions of order n that are in k-sparse or (n,k)-Binomial form as defined in the statement of the theorem, for $k = \lceil 41/\epsilon \rceil$.

The next step is to sparsify this cover by removing elements to obtain the next Theorem 3.3. Note that the term $n \cdot (1/\epsilon)^{O(1/\epsilon^2)}$ in the size of the cover is due to the enumeration over distributions in sparse form. Using Theorem 3.4 below, we argue that there is a lot of redundancy in those distributions, and that it suffices to only include $n \cdot (1/\epsilon)^{O(log^2 1/\epsilon)}$ of them in the cover. In particular, Theorem 3.4 establishes that, if two Poisson Binomial distributions have their first $O(log 1/\epsilon)$ moments equal, then their distance is at most ϵ. So we only need to include at most one sparse form distribution with the same first $O(log 1/\epsilon)$ moments in our cover. We proceed to state Theorem 3.3 including in its proof the Theorem 3.4.

Theorem 3.3. [8]. *For all $n, \epsilon > 0$ there exists a set $S_{n,\epsilon} \subset S_n$ such that:*

1. *$S_{n,\epsilon}$ is an ϵ-cover of S_n in total variation distance; that is, for all $D \in S_n$, there exists some $D' \in S_{n,\epsilon}$ such that $d_{TV}(D, D') \le \epsilon$*

2. *$|S_{n,\epsilon}| \le n^2 + n \cdot (\frac{1}{\epsilon})^{O(log^2 1/\epsilon)}$*

3. *$S_{n,\epsilon}$ can be computed in time $O(n^2 logn) + O(nlogn) \cdot (\frac{1}{\epsilon})^{O(log^2 1/\epsilon)}$*

Moreover, if $\{Y_i\} \in S_{n,\epsilon}$ then the collection of n Bernoulli random variables $\{Y_i\}$, $i = 1, ..., n$ has one of the following forms, where $k = k(\epsilon) = C/\epsilon$ is a positive integer, for some absolute constant $C > 0$:

(a) *(k-sparse form) there exists some $\ell \le k^3$ such that, for all $i \le \ell, \mathbb{E}[Y_i] \in \left\{\frac{1}{k^2}, \frac{2}{k^2}, ..., \frac{k^2-1}{k^2}\right\}$ and, for all $i < \ell, \mathbb{Y}_i \in \{0, 1\}$;or*

(b) *((n, k)-Binomial form) there is some $\ell \in \{1, ..., n\}$ and $q \in \{1, ..., n\}$ such that, for all $i < \ell, \mathbb{E}[Y_i] = q$ and for all $i > \ell, \mathbb{E}[Y_i] = 0$; moreover, ℓ and q satisfy $\ell q \ge k^2$ and $\ell q(1 - q) \ge k^2 - k - 1$.*
Finally, for every $\{X_i\} \in S_n$ for which there is no an ϵ-cover in $S_{n,\epsilon}$ that is in sparse form, there exists some $\{Y_i\} \in S_{n,\epsilon}$ in k-heavy Binomial form such that

(c) *$d_{TV}\left(\sum_{i=1}^n X_i, \sum_{i=1}^n Y_i\right) \le \epsilon$; and*

(d) *If $\mu = \mathbb{E}[\sum_{i=1}^n X_i], \mu' = \mathbb{E}[\sum_{i=1}^n Y_i], \sigma^2 = Var[\sum_{i=1}^n X_i]$and $\sigma'^2 = Var[\sum_{i=1}^n Y_i]$, then$|\mu - \mu'| = O(1)$ and $|\sigma - \sigma'| = O(1 + \epsilon \cdot (1 + \sigma^2))$.*

Proof. Theorem 3.1 implies the existence of an ϵ-cover $S_{n,\epsilon}$ of S_n of size at most $n^2 + n \cdot (\frac{1}{\epsilon})^{O(1/\epsilon^2)}$. This cover is obtained by taking the union of all Poisson Binomial distributions in (n, k)-Binomial form and all Poisson Binomial distributions in k-sparse form, for $k = \lceil 41/\epsilon \rceil$. The total number of Poisson Binomial distributions in (n, k)-Binomial form is at most n^2, since there are at most n choices for the value of ℓ and at most n choices for the value of q. The total number of Poisson Binomial distributions in k-sparse form is at most $(k^3 + 1) \cdot k^{3k^2} \cdot (n + 1) = n \cdot (\frac{1}{\epsilon})^{O(1/\epsilon^2)}$ since there are $k^3 + 1$ choices for ℓ, at most k^{3k^2} choices of probabilities $p_1 \le p_2 \le$

... $\leq p_\ell$ in $\{\frac{1}{k^2}, \frac{2}{k^2}, ..., \frac{k^2-1}{k^2}\}$, and at most $n+1$ choices for the number of variables indexed by $i > \ell$ that have expectation equal to 1. Notice that enumerating over the above distributions takes time $O(n^2 logn) + O(nlogn) \cdot (\frac{1}{\epsilon})^{O(1/\epsilon^2)}$, as a number in $\{0, ..., n\}$ and a probability in $\{\frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n}\}$ can be represented using $O(logn)$ bits, while a number in $\{0, ..., k^3\}$ and a probability in $\{\frac{1}{k^2}, \frac{2}{k^2}, ..., \frac{k^2-1}{k^2}\}$ can be represented using $O(logk) = O(log 1/\epsilon)$ bits.

We next show that we can remove from $S'_{n,\epsilon}$ a large number of the sparse-form distributions it contains to obtain a 2ϵ-cover of S_n. In particular, we shall only keep $n \cdot (\frac{1}{\epsilon})^{O(log^2 1/\epsilon)}$ sparse-form distributions by appealing to the next Theorem.

Theorem 3.4. *Let $\mathcal{P} := (p_i)_{i=1}^n \in [0, 1/2]^n$ and $\mathcal{Q} := (q_i)_{i=1}^n \in [0, 1/2]^n$ be two collections of probability values. Let also $\mathcal{X} := (X_i)_{i=1}^n$ and $\mathcal{Y} := (Y_i)_{i=1}^n$ be two collections of mutually independent indicators with $\mathbb{E}[X_i] = p_i$, for all $i \in [n]$. If for some $d \in [n]$ the following condition is satisfied:*

$$(C_d): \quad \sum_{i=1}^n p_{i=1}^\ell = \sum_{i=1}^n q_{i=1}^\ell, \quad \textit{for all} \quad \ell = 1, ..., d,$$

$$\textit{then} \quad d_{TV}\left(\sum_i X_i, \sum_i Y_i\right) \leq 13(d+1)^{1/4} 2^{-(d+1)/2}.$$

Remark. *Condition (C_d) in the statement of Theorem constrains the first d power sums of the expectations of the constituent indicators of two Poisson Binomial distributions. To relate these power sums to the moments of these distributions we can use the theory of symmetric polynomials to arrive at the following equivalent condition to (C_d) :*

$$(V_d): \quad \mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^\ell\right] = \mathbb{E}\left[\left(\sum_{i=1}^n Y_i\right)^\ell\right], \quad \textit{for all} \quad \lambda \in [d].$$

Theorem 3.4 states that if two sums of independent indicators (equivalent two PBDs) have equal first d moments, then their total variation distance is at most ϵ. Assuming that a class of PBDs in the cover meets the conditions of Theorem 3.4 we will keep only one PBD from the specific class, writing off the others. By repeating the same procedure for all the classes are created we obtain the sparsify cover.

More specific: for a collection $\mathcal{P} = (p_i)_{i\in[n]} \in [0,1]^n$, we denote by $\mathcal{L}_{\mathcal{P}} = \{i|p_i \in (0,1/2]\}$ and by $\mathcal{R}_{\mathcal{P}} = \{i|p_i \in (1/2,1)\}$. For a collection $\mathcal{P} = (p_i)_{i\in[n]} \in [0,1]^n$, we also define its moment profile $m_{\mathcal{P}}$ to be the $(2d(\epsilon)+1)$-dimensional vector

$$m_{\mathcal{P}} = \left(\sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i, \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^2, ..., \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^{d(\epsilon)}; \sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i, ..., \sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i^{d(\epsilon)}; |i|p_i = 1| \right).$$

By Theorem 3.4 if $m_{\mathcal{P}} = m_{\mathcal{Q}}$ then $d_{TV}(PBD(\mathcal{P}), PBD(\mathcal{Q})) \leq \epsilon$. Given the above we will try to sparsify the $S'_{n,\epsilon}$ as follows: for every possible moment profile that can arise from a Poisson Binomial distribution in k-sparse form, we keep in our cover a single Poisson Binomial distribution with such moment profile. The cover resulting from this sparsification is a 2ϵ-cover, since the sparsification loses us an additional ϵ in total variation distance, as argued above.

We now bound the cardinality of the sparsified cover. The total number of moment profiles of k-sparse Poisson Binomial distributions is $k^{O(d(\epsilon)^2)} \cdot (n+1)$. Indeed, consider a Poisson Binomial distribution $PBD(\mathcal{P} = (p_i)_{i\in[n]})$ in k-sparse form. There are at most k^3+1 choices for $|\mathcal{L}_{\mathcal{P}}|$, at most k^3+1 choices for $|\mathcal{R}_{\mathcal{P}}|$, and at most $(n+1)$ choices for $|\{i|p_i = 1\}|$. We also claim that the total number of possible vectors

$$\left(\sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i, \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^2, ..., \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^{d(\epsilon)} \right)$$

is $k^{O(d(\epsilon)^2)}$. Indeed, if $|\mathcal{L}_{\mathcal{P}}| = 0$ there is just one such vector, namely the all-zero vector. If $|\mathcal{L}_{\mathcal{P}}| > 0$, then, for all $t = 1, ..., d(\epsilon)$, $\sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^t \in (0, |\mathcal{L}_{\mathcal{P}}|]$ and it must be an integer multiple of $1/k^{2t}$. So the total number of possible values of $\sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^t$ is at most $k^{2t}|\mathcal{L}_{\mathcal{P}}| \leq k^{2t}k^3$, and the total number of possible vectors

$$\left(\sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i, \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^2, ..., \sum_{i\in\mathcal{L}_{\mathcal{P}}} p_i^{d(\epsilon)}\right)$$

is at most

$$\prod_{t=1}^{d(\epsilon)} k^{2t}k^3 \leq k^{O(d(\epsilon)^2)}.$$

The same upper bound applies to the total number of positive vectors

$$\left(\sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i, \sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i^2, ..., \sum_{i\in\mathcal{R}_{\mathcal{P}}} p_i^{d(\epsilon)}\right).$$

The moment profiles we enumerated over are a superset of the moment profiles of k-sparse Poisson Binomial distributions. We call them compatible moment profiles. We argued that there are at most $k^{O(d(\epsilon)^2)} \cdot (n+1)$ compatible moment profiles, so the total number of Poisson Binomial distributions in k-sparse form that we keep in the cover is at most $k^{O(d(\epsilon)^2)} \cdot (n+1) = n \cdot (\frac{1}{\epsilon})^{O(log^2/\epsilon)}$. The number of Poisson Binomial distributions in (n,k)-Binomial form is the same as before, i.e. at most n^2 , as we did not eliminate any of them. So the size of the sparsified cover is $n^2 + n \cdot (\frac{1}{\epsilon})^{O(log^2/\epsilon)}$.

To finish the proof it remains to argue that we don't actually need to first compute $S_{n,\epsilon}$ and then sparsify it to obtain our cover, but can produce it directly in time $O(n^2 logn) + O(nlogn) \cdot (\frac{1}{\epsilon})^{O(log^2/\epsilon)}$. We claim that, given a moment profile m that is compatible with a k-sparse Poisson Binomial distribution, we can compute some $PBD(\mathcal{P} = (p_i)_i)$ in k-sparse form such that $m_{\mathcal{P}} = m$, if such a distribution exists, in time $O(logn) \cdot (\frac{1}{\epsilon})^{O(log^2/\epsilon)}$. So the algorithm enumerates over all moment profiles that are compatible with a k-sparse Poisson Binomial distribution and for each profile finds a Poisson Binomial distribution with such moment profile, if such distribution exists, adding it to the cover if it does exist. It then enumerates over all Poisson Binomial distributions in (n,k)-Binomial form and adds them to the cover as well. The overall running time is as promised.

Chapter 4

A Learning Algorithm

In this chapter we describe the main result of [8] regarding an efficient algorithm for learning PBDs from $O(1/\epsilon^2)$ many samples independent of $[n]$. Since PBDs are an n-parameter family of distributions over the domain $[n]$, the view of such a tight bound is a surprising result. The starting point of the algorithm for learning PBDs is a theorem of [41, 42] that gives detailed information about the structure of a small ϵ-cover (under the total variation distance) of the space of all PBDs on n variables (see Theorem 3.3). Roughly speaking, this result says that every PBD is either close to a PBD whose support is sparse, or is close to a translated "heavy" Binomial distribution. The learning algorithm exploits this structure of the cover; it has two subroutines corresponding to these two different types of distributions that the cover contains. First, assuming that the target PBD is close to a sparsely supported distribution, it runs Birge's unimodal distribution learner over a carefully selected subinterval of $[n]$ to construct a hypothesis H_S; the (purported) sparsity of the distribution makes it possible for this algorithm to use $O(1/\epsilon^3)$ samples independent of n. Then, assuming that the target PBD is close to a translated "heavy" Binomial distribution, the algorithm constructs a hypothesis Translated Poisson Distribution H_P [43] whose mean and variance match the estimated mean and variance of the target PBD; the

H_P is close to the target PBD if the target PBD is not close to any sparse distribution in the cover. At this point the algorithm has two hypothesis distributions, H_S and H_P, one of which should be good; it remains to select one as the final output hypothesis. This is achieved using a form of "hypothesis testing" for probability distributions.

4.1 Learning Poisson Binomial Distributions

Theorem 4.1. [8]. *Let $X = \sum_{i=1}^{n} X_i$, be an unknown PBD.*

1. ***[Learning PBDs from constantly many samples]*** *There is an algorithm with the following properties: given n, ϵ, δ and access to independent draws from X, the algorithm uses*

   $$O((1/\epsilon^3) \cdot log(1/\delta))$$

   samples from X, performs

   $$O((1/\epsilon^3) \cdot logn \cdot log^2(1/\delta))$$

   bit operations, and with probability at least $1-\delta$ outputs a (succinct description of a) distribution $\hat{X}$ over $[n]$ which is such that $d_{TV}(X, \hat{X}) \leq \epsilon$.

2. ***[Properly learning PBDs from constantly many samples]*** *There is an algorithm with the following properties: given n, ϵ, δ and access to independent draws from X, the algorithm uses*

   $$O((1/\epsilon^2) \cdot log(1/\delta))$$

   samples from X, performs

   $$(1/\epsilon)^{O(log^2(1/\epsilon))} \cdot O(logn \cdot log1/\delta)$$

   bit operations, and with probability at least $1 - \delta$ outputs a (succinct description of a) vector $\hat{p} = (\hat{p}_1, ..., \hat{p}_n)$ defining a PBD $\hat{X}$ such that $d_{TV}(X, \hat{X}) \leq \epsilon$.

The Basic Learning Algorithm. The high-level structure of the learning algorithms which give theorem 4.1 is provided in Algorithm $Learn{-}PBD$ of Figure 1

```
$Learn - PBD(n, \epsilon, \delta)$
    1. Run $Learn - Sparse^X(n, \epsilon, \delta/3)$ to get hypothesis distribution $H_S$.
    2. Run $Learn - Poisson^X(n, \epsilon, \delta/3)$ to get hypothesis distribution $H_P$.
    3. Return the distribution which is the output of
       $Choose - Hypothesis^X(H_S, H_P, \epsilon, \delta/3)$.
```

Figure 1: Learn-PBD(n, ϵ, δ)

At a high level, the subroutine $Learn - Sparse$ is given sample access to X and is designed to find an ϵ-accurate hypothesis H_S with probability at least $1 - \delta/3$, if the unknown PBD X is ϵ-close to some sparse form PBD inside the cover $S_{n,\epsilon}$. Similarly, $Learn - Poisson$ is designed to find an ϵ-accurate hypothesis H_P, if X is not ϵ-close to a sparse form PBD (in this case, Theorem 3.3 implies that X must be ϵ-close to some $k(\epsilon)$-heavy Binomial form PBD). Finally, $Choose - Hypothesis$ is designed to choose one of the two hypothesis H_S, H_P as being ϵ-close to X. The following subsections specify these subroutines, as well as how the algorithm can be used to establish Theorem 3.3. Note that $Learn - Sparse$ and $Learn - Poisson$ do not return the distributions H_S and H_P as a list of probabilities for every point in $[n]$. They return instead a succinct description of these distributions in order to keep the running time of the algorithm logarithmic in n. Similarly, $Choose - Hypothesis$ operates with succinct descriptions of these distributions.

4.1.1 Learning when X is close to a sparse PBD

The starting point here is the simple observation that any PBD is a unimodal distribution over the domain $\{0, 1, ..., n\}$. (There is a simple inductive proof of this, or see Section 2 of [44].) This enables us to use the algorithm of Birge [4] for learning unimodal distributions. The Theorem 4.2 as stated below follows from [4].

Theorem 4.2. [4]. *For all $n, \epsilon, \delta > 0$, there is an algorithm that draws*

$$O(\frac{logn}{\epsilon^3}log\frac{1}{\delta} + \frac{1}{\epsilon^2}log\frac{1}{\delta}loglog\frac{1}{\delta})$$

samples from an unknown unimodal distribution X over $[n]$, does

$$O(\frac{log^2 n}{\epsilon^2}log^2\frac{1}{\delta})$$

bit-operations, and outputs a (succinct description of a) hypothesis distribution H over $[n]$ that has the following form: H is uniform over subintervals $[\alpha_1, \beta_1], [\alpha_2, \beta_2], ..., [\alpha_k, \beta_k]$, whose union $\bigcup_{i=1}^{n}[\alpha_i, \beta_i] = [n]$, where $k = O(\frac{logn}{\epsilon})$. In particular, the algorithm outputs the lists α_1 through α_k and β_1 through β_k, as well as the total probability mass that H assigns to each subinterval $[\alpha_i, \beta_i], i = 1, ..., k$. Finally, with probability at least $1-\delta, d_{TV}(X, H) \leq \epsilon$.

The main result of this subsection is the following:

Lemma 4.3. *For all $n, \epsilon', \delta' > 0$, there is an algorithm $Learn{-}Sparse^X(n, \epsilon', \delta')$ that draws*

$$O(\frac{1}{\epsilon'^3}log\frac{1}{\epsilon'}log\frac{1}{\delta'} + \frac{1}{\epsilon'^2}log\frac{1}{\delta'}loglog\frac{1}{\delta'})$$

samples from a target PBD X over $[n]$, does

$$logn \cdot O(\frac{1}{\epsilon'^3}log^2\frac{1}{\delta'})$$

bit operations, and outputs a (succinct description of a) hypothesis distribution H_S over $[n]$ that has the following form: its support is contained in an *explicitly specified interval* $[a, b] \subset [n]$*, where* $|b - a| = O(\frac{1}{\epsilon'^3})$*, and for every point in* $[a, b]$ *the algorithm explicitly specifies the probability assigned to that point by* H_S*. The algorithm has the following guarantee: if* X *is* ϵ'*-close to some sparse form PBD* Y *in the cover* $S_{n,\epsilon'}$ *of Theorem 3.3, then with probability at least* $1 - \delta', d_{TV}(X, H_S) \leq c_1\epsilon'$*, for some absolute constant* $c_1 \geq 1$*, and the support of* H_S *lies in the support of* Y*.*

The high-level idea of Lemma 4.3 is quite simple. Truncate $O(\epsilon')$ of the probability mass from each end of X to obtain a conditional distribution $X_{[\hat{a},\hat{b}]}$; since X is unimodal so is $X_{[\hat{a},\hat{b}]}$. If $\hat{b} - \hat{a}$ is larger than $O(1/\epsilon^{'3})$ then the algorithm outputs "fail" (and X could not have been close to a sparse-form distribution in the cover). Otherwise, use Birge's algorithm to learn the unimodal distribution $X_{[\hat{a},\hat{b}]}$. A detailed description of the algorithm is given in Figure 2 below.

$Learn - Sparse^X(n, \epsilon', \delta')$

1. Draw $M = 32log(8/\delta')\epsilon^{'2}$ samples from X and sort them to obtain a list of values $0 \leq s_1 \leq ... \leq s_M \leq n$.
2. Define $\hat{a} := s_{\lceil 2\epsilon' M \rceil}$ and $\hat{b} := s_{\lfloor (1-2\epsilon')M \rfloor}$.
3. If $\hat{b} - \hat{a} > (C/\epsilon')^3$ (where C is the constant in the statement of Theorem 3.3), output "fail" and return the (trivial) hypothesis which puts probability mass 1 on the point 0.
4. Otherwise, run Birge's unimodal distribution learner (Theorem 4.2) on the conditional distribution $X_{[\hat{a},\hat{b}]}$ and output the hypothesis that it returns.

Figure 2: $Learn - Sparse^X n, \epsilon', \delta'$

Proof. As described in Figure 2, algorithm $Learn - Sparse^X n, \epsilon', \delta'$ first draws $M = 32log(8/\delta')\epsilon^{'2}$ samples from X and sorts them to obtain a list of values $0 \leq s_1 \leq ... \leq s_M \leq n$. The following claim holds about the values $\hat{a}$ and $\hat{b}$ defined in Step 2 of the algorithm:

Claim 1. With probability at least $1-\delta'/2$, we have $Pr[X \leq \hat{a}] \in [3\epsilon'/2, 5\epsilon'/2]$ and $Pr[X \leq \hat{b}] \in [1-5\epsilon'/2, 1-3\epsilon'/2]$.

Proof of Claim. We only show that $Pr[X \leq \hat{a}] \geq 3\epsilon'/2$ with probability at least $1-\delta'/8$, since the arguments for $Pr[X \leq \hat{a}] \leq 5\epsilon'/2, Pr[X \leq \hat{b}] \leq 1-3\epsilon'/2$ and $Pr[X \leq \hat{b}] \geq 1-5\epsilon'/2$ are identical. Given that each of these conditions is met with probability at least $1-\delta'/8$, the union bound establishes the claim. To show that $Pr[X \leq \hat{a}] \geq 3\epsilon'/2$ is satisfied with probability at least $1-\delta'/8$ we argue as follows: Let $a' = max\{i|Pr[X] < 3\epsilon'/2\}$. Clearly, $Pr[X \leq a'] < 3\epsilon'/2$ while $Pr[X \leq a'+1] \geq 3\epsilon'/2$. Given this, if M samples are drawn from X then the expected number of them that are $\leq a'$ is at most $3\epsilon' M/2$. It follows then from the $Chernoff$ bound that the probability that more than $7/4\epsilon' M$ samples are $\leq a'$ is at most $e^{-(\epsilon'/4)^2 M/2} \leq \delta'/8$. Hence except with this failure probability, we have $\hat{a} \geq a'+1$, which implies that $Pr[X \leq \hat{a}] \geq 3\epsilon'/2$.

As specified in Steps 3 and 4, if $\hat{b}-\hat{a} > (C/\epsilon')^3$, where C is the constant in the statement of Theorem 3.3, the algorithm outputs “fail”, returning the trivial hypothesis which puts probability mass 1 on the point 0. Otherwise, the algorithm runs Birge's unimodal distribution learner (Theorem 4.2) on the conditional distribution $X_{[\hat{a},\hat{b}]}$, and outputs the result of Birge's algorithm. Since X is unimodal, it follows that $X_{[\hat{a},\hat{b}]}$ is also unimodal, hence Birge's algorithm is appropriate for learning it. The way the Birge's algorithm is applied to learn $X_{[\hat{a},\hat{b}]}$ given samples from the original distribution X is the obvious one: draw samples from X, ignoring all samples that fall outside of $[\hat{a},\hat{b}]$, until the right $O(log(1/\delta')log(1/\epsilon')/\epsilon'^3$ number of samples fall inside $[\hat{a},\hat{b}]$, as required by Birge's algorithm for learning a distribution of support of size $(C/\epsilon')^3$ with probability at least $1-\delta'/4$. Once the right number of samples in $[\hat{a},\hat{b}]$ has been obtained, the algorithm runs Birge's algorithm to learn the conditional distribution $X_{[\hat{a},\hat{b}]}$. Note that the number of samples we need to draw from X until the right $O(log(1/\delta')log(1/\epsilon')/\epsilon^{'3}$ number of samples fall inside $[\hat{a}, \hat{b}]$ is still $O(log(1/\delta')log(1/\epsilon')/\epsilon^{'3}$, with probability at least $1-\delta'/4$. Indeed, since $P(\hat{a} \leq X \leq \hat{b}) = 1-O(\epsilon')$, it follows from the $Chernoff$ bound that with probability at least $1-\delta'/4$, if $K = \Theta(log(1/\delta')log(1/\epsilon')/\epsilon^{'3})$ samples are drawn from X, at least $K(1-O(\epsilon'))$ fall inside $[\hat{a}, \hat{b}]$.

Analysis: It is easy to see that the sample complexity of our algorithm is as promised. For the running time, notice that, if Birge's algorithm is invoked, it will return two lists of numbers a_1 through a_k and b_1 through b_k, as well as a list of probability masses $q_1, ..., q_k$ assigned to each subinterval $[a_i, b_i], i = 1, ..., k$, by the hypothesis distribution H_S, where $k = O(log(1/\epsilon')/\epsilon')$. In linear time, we can compute a list of probabilities $\hat{q}_1, ..., \hat{q}_k$, representing the probability assigned by H_S to every point of subinterval $[a_i, b_i]$, for $i = 1, ..., k$. So we can represent our output hypothesis H_S via a data structure that maintains $O(1/\epsilon'^3)$ pointers, having one pointer per point inside $[a, b]$. The pointers map points to probabilities assigned by H_S to these points. Thus turning the output of Birge's algorithm into an explicit distribution over $[a, b]$ incurs linear overhead in our running time, and hence the running time of our algorithm is also as promised. Moreover, we also note that the output distribution has the promised structure, since in one case it has a single atom at 0 and in the other case it is the output of Birge's algorithm on a distribution of support of size $(C/\epsilon')^3$.

It only remains to justify the last part of the lemma. Let Y be the sparse-form PBD that X is close to; say that Y is supported on $\{a', ..., b'\}$ where $b' - a' \leq (C/\epsilon')^3$. Since X is ϵ' -close to Y in total variation distance it must be the case that $P[X \leq a' - 1] \leq \epsilon'$. Since $P[X \leq \hat{a}' - 1] \geq 3\epsilon'/2$ by Claim 1, it must be the case that $\hat{a} \geq a'$. Similar arguments give that $\hat{b} \leq b'$. So the interval $[\hat{a}, \hat{b}]$ is contained in $[a', b']$ and has length at most $(C/\epsilon')^3$. This means that Birge's algorithm is indeed used correctly by our algorithm to learn $X_{[\hat{a}, \hat{b}]}$, with probability at least $1 - \delta'/2$ (that is, unless Claim 1 fails). Now it follows from the correctness of Birge's algorithm (Theorem 4.2) and the discussion above, that the hypothesis H_S output when Birge's algorithm is invoked satisfies $d_{TV}(H_S, X_{[\hat{a},\hat{b}]}) \leq \epsilon'$, with probability at least $1 - \delta'/2$, i.e., unless either Birge's algorithm fails, or we fail to get the right number of samples landing inside $[\hat{a}, \hat{b}]$. To conclude the proof of the lemma we note that:

$$2d_{TV}(X, X_{[\hat{a},\hat{b}]}) = \sum_{i\in[\hat{a},\hat{b}]} |Pr[X_{[\hat{a},\hat{b}]} = i] - Pr[X = i]| + \sum_{i\notin[\hat{a},\hat{b}]} |Pr[X_{[\hat{a},\hat{b}]} = i] - Pr[X = i]|$$

$$= \sum_{i\in[\hat{a},\hat{b}]} \left| \frac{1}{\sum_{i\in[\hat{a},\hat{b}]} Pr[X = i]} Pr[X = i] - Pr[X = i] \right| + \sum_{i\notin[\hat{a},\hat{b}]} Pr[X = i]$$

$$= \sum_{i\in[\hat{a},\hat{b}]} \left| \frac{1}{1 - O(\epsilon')} Pr[X = i] - Pr[X = i] \right| + O(\epsilon')$$

$$\frac{O(\epsilon')}{1 - O(\epsilon')} \sum_{i\in[\hat{a},\hat{b}]} |Pr[X = i]| + O(\epsilon') = O(\epsilon')$$

So the triangle inequality gives: $d_{TV}(H_S, X) = O(\epsilon')$ and the lemma is proved.

4.1.2 Learning when X is close to a k-heavy Binomial Form PBD

Lemma 4.4. *For all $n, \epsilon', \delta' > 0$, there is an algorithm $Learn{-}Poisson^X(n, \epsilon', \delta')$ that draws $O(log(1/\delta')/\epsilon'^2)$ samples from a target PBD X over $[n]$, does $O(logn \cdot log(1/\delta')/\epsilon^2)$ bit operations, and returns two parameters $\hat{\mu}$ and $\hat{\sigma}^2$. The algorithm has the following guarantee: Suppose X is not ϵ'-close to any sparse form PBD in the cover $S_{n,\epsilon'}$ of Theorem 3.3. Let $H_P = TP(\hat{\mu}, \hat{\sigma}^2)$ be the translated Poisson distribution with parameters $\hat{\mu}$ and $\hat{\sigma}^2$. Then with probability at least $1 - \delta'$ we have $d_{TV}(X, H_P) \leq c_2\epsilon'$ for some absolute constant $c_2 \geq 1$.*

Our proof plan is to exploit the structure of the cover of Theorem 3.3. In particular, if X is not ϵ'-close to any sparse form PBD in the cover, it must be ϵ'-close to a PBD in heavy Binomial form with approximately the same mean and variance as X, as specified by the final part of the cover theorem. Hence, a natural strategy is to obtain estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of the mean and variance of the unknown PBD X, and output as a hypothesis a translated Poisson distribution with parameters $\hat{\mu}$ and $\hat{\sigma}^2$. This strategy is a successful one. Before providing the details, two facts should be highlighted as there will be used later. The first is that, assuming X is not ϵ'-close to any sparse form PBD in the cover $S_{n,\epsilon'}$, its variance σ^2 satisfies

$$\sigma^2 = \Omega(1/\epsilon^{'2}) \geq \theta^2 \quad \text{for some universal constant} \quad \theta. \tag{4.1}$$

The second is that under the same assumption, the estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of the mean μ and variance σ^2 of X that we obtain satisfy the following bounds with probability at least $1-\delta$:

$$|\mu - \hat{\mu}| \leq \epsilon' \cdot \sigma \quad \text{and} \quad |\sigma^2 - \hat{\sigma}^2| \leq \epsilon' \cdot \sigma^2. \tag{4.2}$$

```
$Learn - Poisson^X(n, \epsilon', \delta')$
    1. Let $\epsilon = \epsilon'/\sqrt{4 + \frac{1}{\theta^2}}$ and $\delta = \delta'$.
    2. Run algorithm $A(n, \epsilon, \delta)$ to obtain an estimate $\hat{\mu}$ of $\mathbb{E}[X]$ and an estimate $\hat{\sigma}^2$ of $Var[X]$.
    3. Output the translated Poisson distribution $TP(\hat{\mu}, \hat{\sigma}^2)$.
```

Figure 3: $Learn - Poisson^X(n, \epsilon', \delta')$ The value θ used in Line 1 is the universal constant specified in the proof of Lemma 4.4

```
$A(n, \epsilon, \delta)$

1. Let $r = O(log 1/\delta)$. For $i = 1, ..., r$ repeat the following:
    (a) Draw $m = \lceil 3/\epsilon^2 \rceil$ independent samples $Z_{i,1}, ..., Z_{i,m}$ from $X$.
    (b) Let $\hat{\mu}_i = \frac{\sum_j Z_{i,j}}{m}, \hat{\sigma}_i^2 = \frac{\sum_j \left(Z_{i,j} - \frac{1}{m}\sum_k Z_{i,k}\right)^2}{m-1}$.
2. Set $\hat{\mu}$ to be the median of $\hat{\mu}_1, ..., \hat{\mu}_r$ and set $\hat{\sigma}^2$ to be the median of $\hat{\sigma}_1^2, ..., \hat{\sigma}_r^2$.
3. Output $\hat{\mu}$ and $\hat{\sigma}^2$.
```

Figure 4: $A(n, \epsilon, \delta)$

Lemma 4.5. *For all $n, \epsilon, \delta > 0$, there exists an algorithm $A(n, \epsilon, \delta)$ with the following properties: given access to a PBD X of order n, it produce estimates $\hat{\mu}$ and $\hat{\sigma}^2$ for $\mu = \mathbb{E}[X]$ and $\sigma^2 = Var[X]$ respectively such that with probability at least $1 - \delta$:*

$$|\mu - \hat{\mu}| \leq \epsilon \cdot \sigma \quad \textit{and} \quad |\sigma^2 - \hat{\sigma}^2| \leq \epsilon \cdot \sigma^2 \sqrt{4 + \frac{1}{\sigma^2}}.$$

The algorithm uses

$$O(log(1/\delta)/\epsilon^2)$$

samples and runs in time

$$O(logn \cdot log(1/\delta)/\epsilon^2).$$

Proof. We treat the estimation of μ and σ^2 separately. For both estimation problems we show how to use $O(1/\epsilon^2)$ samples to obtain estimates $\hat{\mu}$ and $\hat{\sigma}^2$ achieving the required guarantees with probability at least 2/3 (we refer to these as "weak estimators"). Then a routine procedure allows us to boost the success probability to $1 = \delta$ at the expense of a multiplicative factor $O(log 1/\delta)$ on the number of samples. While we omit the details of the routine boosting argument, we remind the reader that it involves running the weak estimator $O(log 1/\delta)$ times to obtain estimates $\hat{\mu}_1, ..., \hat{\mu}_{O(log 1/\delta)}$ and outputting the median of these estimates, and similarly for estimating σ^2. We proceed to specify and analyze the weak estimators for μ and σ^2 separately:

Weak estimator for μ.

Let $Z_1, ..., Z_m$ be independent samples from X, and let $\hat{\mu} = \frac{1}{m}\sum_i Z_i$. Then

$$\mathbb{E}[\hat{\mu}] = \mu \quad \text{and} \quad Var[\hat{\mu}] = \frac{1}{m}\sigma^2.$$

Chebyshev's inequality implies that

$$Pr\left[|\mu - \hat{\mu}| \geq \frac{t\sigma}{\sqrt{m}}\right] \leq \frac{1}{t^2}$$

Choosing $t = \sqrt{3}$ and $m = \lceil \frac{3}{\epsilon^2} \rceil$, the above inequality implies that $|\mu - \hat{\mu}| \leq \epsilon \cdot \sigma$ with probability at least $\frac{2}{3}$.

Weak estimator for σ^2.

Let $Z_1, ..., Z_m$ be independent samples from X, and let $\hat{\sigma}^2 = \frac{\sum_i (Z_i - \frac{1}{m}\sum_i Z_1)^2}{m-1}$ be the unbiased sample variance. Then

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \quad \text{and} \quad Var[\hat{\sigma}^2] = \sigma^4\left(\frac{2}{m-1} + \frac{k}{m}\right).$$

where k is the excess kurtosis of the distribution of X. To bound k in terms of σ^2 suppose that $X = \sum_{i=1}^n X_i$, where $\mathbb{E}[X_i] = p_i$ for all i. Then

$$\begin{aligned} k &= \frac{1}{\sigma^4}\sum_i (1 - 6p_i(1-p_i))(1-p_i)p_i \quad \text{(see [51])} \\ &\leq \frac{1}{\sigma^4}\sum_i (1-p_i)p_i \\ &= \frac{1}{\sigma^2}. \end{aligned}$$

Hence $Var[\hat{\sigma}^2] = \sigma^4\left(\frac{2}{m-1} + \frac{k}{m}\right) \leq \frac{\sigma^4}{m}(4 + \frac{1}{\sigma^2})$. So Chebyshev's inequality implies that

$$Pr\left[|\hat{\sigma}^2 - \sigma^2| \geq t\frac{\sigma^2}{\sqrt{m}}\sqrt{4 + \frac{1}{\sigma^2}}\right] \leq \frac{1}{t^2}.$$

Choosing $t = \sqrt{3}$ and $m = \lceil \frac{3}{\epsilon^2} \rceil$ the above imply that $|\hat{\sigma}^2 - \sigma^2| \leq \epsilon \cdot \sigma^2 \sqrt{4 + \frac{1}{\sigma^2}}$ with probability at least $\frac{2}{3}$.

We proceed to prove Lemma 4.4. $Learn - Poisson^X(n, \epsilon', \delta')$ runs $A(n, \epsilon, \delta)$ from Lemma 4.5 with appropriately chosen $\epsilon = \epsilon(\epsilon')$ and $\delta = \delta(\delta')$, given below, and then outputs the translated Poisson distribution $TP(\hat{\mu}, \hat{\sigma}^2)$, where $\hat{\mu}$ and $\hat{\sigma}^2$ are the estimated mean and variance of X output by A. Next, we show how to choose ϵ and δ, as well as why the desired guarantees are satisfied by the output distribution.

If X is not ϵ'-close to any PBD in sparse form inside the cover $S_{\epsilon'}$ of Theorem 3.1, there exists a PBD Z in $(k = O(1/\epsilon'))$-heavy Binomial form inside S'_ϵ that is within total variation distance ϵ' from X. We use the existence of such Z to obtain lower bounds on the mean and variance of X. Indeed, suppose that the distribution of Z is $Bin(\ell, q)$, a Binomial with parameters ℓ, q. Then Theorem 3.1 certifies that the following conditions are satisfied by the parameters $\ell, q, \mu = \mathbb{E}[X]$ and $\sigma^2 = Var[X]$:

1. $\ell q \geq k^2$
2. $\ell q(1-q) \geq k^2 - k - 1$
3. $|\ell q - \mu| = O(1)$ and
4. $|\ell q(1-q) - \sigma^2| = O(1 + \epsilon' \cdot (1 + \sigma^2))$.

In particular, conditions (2) and (3) above imply that

$$\sigma^2 = \Omega(k^2) = \Omega(1/\epsilon'^2) \geq \theta^2,$$

for some universal constant θ, establishing (4.1). In terms of this θ, we choose $\epsilon = \epsilon' / \sqrt{4 + \frac{1}{\theta^2}}$ and $\delta = \delta'$ for the application of lemma 2.6 to obtain estimates $\hat{\mu}$ and $\hat{\sigma}^2$.

From the choice of parameters and the guarantees of lemma 2.6, it follows that, if X is not ϵ'-close to any PBD in sparse form inside the cover $S_{\epsilon'}$, then with probability at least $1-\delta'$ the estimates $\hat{\mu}$ and $\hat{\sigma}^2$ satisfy:

$$|\mu-\hat{\mu}| \le \epsilon' \cdot \sigma \quad \text{and} \quad |\sigma^2-\hat{\sigma}^2| \le \epsilon' \cdot \sigma^2,$$

establishing (4.2). Moreover, if Y is a random variable distributed according to the translated Poisson distribution $TP(\hat{\mu},\hat{\sigma}^2)$, we show that X and Y are within $O(\epsilon')$ in total variation distance, concluding the proof of lemma 4.4.

Claim 2: If X and Y are as above, then $d_{TV}(X,Y) \le O(\epsilon')$.

Proof. Suppose that $X = \sum_{i=1}^n X_i$, where $\mathbb{E}[X_i] = p_i$ for all i. Lemma 2.6 implies that

$$d_{TV}\left(\sum_i X, TP(\mu,\sigma^2)\right) \le \frac{\sqrt{\sum_i p_i^3(1-p_i)}+2}{\sum_i p_i(1-p_i)} \le \frac{\sqrt{\sum_i p_i(1-p_i)}+2}{\sum_i p_i(1-p_i)}$$

$$\le \frac{1}{\sqrt{\sum_i p_i(1-p_i)}} + \frac{2}{\sum_i p_i(1-p_i)} = \frac{1}{\sigma}+\frac{2}{\sigma^2}$$

$$= O(\epsilon').$$

It remains to bound the total variation distance between the translated Poisson distributions $TP(\mu,\sigma^2)$ and $TP(\hat{\mu},\hat{\sigma}^2)$, by lemma 2.4

$$d_{TV}(TP(\mu,\sigma^2), TP(\hat{\mu},\hat{\sigma}^2)) \le \frac{|\mu-\hat{\mu}|}{min(\sigma,\hat{\sigma})} + \frac{|\sigma^2-\hat{\sigma}^2|+1}{min(\sigma^2,\hat{\sigma}^2)}$$

$$\le \frac{\epsilon'\sigma}{min(\sigma,\hat{\sigma})} + \frac{\epsilon'\sigma^2+1}{min(\sigma^2,\hat{\sigma}^2)}$$

$$\le \frac{\epsilon'\sigma}{\sigma/\sqrt{1-\epsilon'}} + \frac{\epsilon'\sigma^2+1}{\sigma^2/\sqrt{1-\epsilon'}}$$

$$= O(\epsilon') + \frac{O(1-\epsilon')}{\sigma^2}$$

$$= O(\epsilon') + O(\epsilon'^2)$$

$$= O(\epsilon').$$

The claim follows from the triangle inequality

$$d_{TV}(X, Y)$$

$$\leq d_{TV}\left(\sum_i X, TP(\mu, \sigma^2)\right) + d_{TV}(TP(\mu, \sigma^2), TP(\hat{\mu}, \hat{\sigma}^2)) + d_{TV}(TP\left(\hat{\mu}, \hat{\sigma}^2), \sum_i Y\right)$$

$$\leq 3O(\epsilon') = O(\epsilon').$$

The proof of Lemma 4.4 is concluded. We remark that the algorithm described above does not need to know a priori whether or not X is ϵ'-close to a PBD in sparse form inside the cover $S_{\epsilon'}$ of Theorem 3.1. The algorithm simply runs the estimator of Lemma 4.5 with $\epsilon = \epsilon'/\sqrt{4 + \frac{1}{\theta^2}}$ and $\delta = \delta'$ and outputs whatever estimates $\hat{\mu}$ and $\hat{\sigma}^2$ the algorithm of Lemma 4.5 produces.

4.1.3 Hypothesis testing

The hypothesis testing routine $Choose - Hypothesis^X$ uses samples from the unknown distribution X to run a "competition" between two candidate hypothesis distributions H_1 and H_2 over $[n]$ that are given in the input. It is proven that if at least one of the two candidate hypotheses is close to the unknown distribution X, then with high probability over the samples drawn from X the routine selects as winner a candidate that is close to X. This basic approach of running a competition between candidate hypotheses is quite similar to the "Scheffe estimate" proposed by Devroye and Lugosi (see [45, 46] and Chapter 6 of [47], as well as [48]), but the notion of competition here is different.

Lemma 4.6. . *There is an algorithm* $Choose - Hypothesis^X(H_1, H_2, \epsilon', \delta')$ *which is given sample access to distribution* X*, two hypothesis distributions*

H_1, H_2 for X, an accuracy parameter $\epsilon' > 0$, and a confidence parameter $\delta' > 0$. It makes

$$m = O(log(1/\delta')/\epsilon^{'2})$$

draws from X and returns some $H \in \{H_1, H_2\}$. If $d_{TV}(X, H_i) \leq \epsilon'$ for some $i \in \{1, 2\}$, then with probability at least $1 - \delta'$ the distribution H that $Choose - Hypothesis$ returns has $d_{TV}(X, H) \leq 6\epsilon'$.

Proof of Lemma 4.6: Figure 5 describes how the competition between H_1 and H_2 is carried out.

$Choose - Hypothesis^X(H_1, H_2, \epsilon', \delta')$

INPUT: Sample access to distribution X; a pair of hypothesis distributions (H_1, H_2); $\epsilon', \delta' > 0$.

Let W be the support of $X, W_1 = W_1(H_1, H_2) := \{w \in W | H_1(W) > H_2(W)\}$, and $p_1 = H_1(W_1), p_2 = H_2(W_1)$. /* Clearly, $p_1 > p_2$ and $d_{TV}(H_1, H_2) = p_1 - p_2$.*/

1. If $p_1 - p_2 \leq 5\epsilon'$, declare a draw and return either H_i. Otherwise:
2. Draw $m = 2log(1/\delta')/\epsilon^{'2}$ samples $s_1, ..., s_m$ from X, and let $\tau = \frac{1}{m}|\{i | s_i \in W_1\}|$ be the fraction of samples that fall inside W_1.
3. If $\tau > p_1 - \frac{3}{2}\epsilon'$, declare H_1 as winner and return H_1; otherwise,
4. if $\tau < p_2 + \frac{3}{2}\epsilon'$, declare H_2 as winner and return H_2; otherwise,
5. declare a draw and return either H_i.

Figure 5: $Choose - Hypothesis^X(H_1, H_2, \epsilon', \delta')$

The correctness of $Choose - Hypothesis$ is an immediate consequence of the following claim. (In fact for Lemma 4.6 we only need item (i) below, but item (ii) will be handy later in the proof of Lemma 4.7.)

Claim 3. Suppose that $d_{TV}(X, H_i) \leq \epsilon'$ for some $i \in \{1, 2\}$. Then:

1. if $d_{TV}(X, H_{3-i}) > 6\epsilon'$, the probability that $Choose{-}Hypothesis^X(H_1, H_2, \epsilon', \delta')$ does not declare H_i as the winner is at most $2e^{-m\epsilon'^2/2}$, where m is chosen as in the description of the algorithm. (Intuitively, if H_{3-i} is very bad then it is very likely that H_i will be declared winner.)

2. if $d_{TV}(X, H_{3-i}) > 4\epsilon'$, the probability that $Choose{-}Hypothesis^X(H_1, H_2, \epsilon', \delta')$ declares H_{3-i} as the winner is at most $2e^{-m\epsilon'^2/2}$. (Intuitively, if H_{3-i} is only moderately bad then a draw is possible but it is very unlikely that H_{3-i} will be declared winner.)

Proof. Let $r = X(W_1)$. The definition of the total variation distance implies that $|r - p_i| \leq \epsilon'$. Let us define independent indicators $\{Z_j\}_{j=1}^m$ such that, for all j, $Z_j = 1$ iff $s_j \in W_1$. Clearly, $\tau = \frac{1}{m}\sum_{j=1}^m Z_j$ and $\mathbb{E}[\tau] = \mathbb{E}[Z_j] = r$. Since the Z_j's are mutually independent, it follows from the $Chernoff bound$ that $Pr[|\tau - r| \geq \epsilon'/2] \leq 2e^{-m\epsilon'^2/2}$. Using $|r - p_i| \leq \epsilon'$ we get that $Pr[|\tau - p_i|] \geq 3\epsilon'/2 \leq 2e^{-m\epsilon'^2/2}$. Hence:

- For part (i): If $d_{TV}(X, H_{3-i}) > 6\epsilon'$, from the triangle inequality we get that $p_1 - p_2 = d_{TV}(H_1, H_2) > 5\epsilon'$. Hence, the algorithm will go beyond step 1, and with probability at least $1 - 2e^{-m\epsilon'^{2/2}}$, it will stop at step 3 (when $i = 1$) or step 4 (when $i = 2$), declaring H_i as the winner of the competition between H_1 and H_2.

- For part (ii): If $p_1 - p_2 \leq 5\epsilon'$ then the competition declares a draw, hence H_{3-i} is not the winner. Otherwise we have $p_1 - p_2 > 5\epsilon'$ and the above arguments imply that the competition between H_1 and H_2 will declare H_{3-i} as the winner with probability at most $2e^{-m\epsilon'^{2/2}}$.

This concludes the proof of Claim 3. In view of Claim 3, the proof of Lemma 4.6 is concluded. $Choose - Hypothesis$ algorithm implies a generic learning algorithm of independent interest.

Lemma 4.7. . *Let S be an arbitrary set of distributions over a finite domain. Moreover, let $S_{n,\epsilon} \subseteq S$ be an ϵ-cover of S of size N, for some $\epsilon > 0$. For all*

$\delta > 0$, there is an algorithm that uses

$$O(\epsilon^{-2} logN log(1/\delta))$$

samples from an unknown distribution $X \in S$ and, with probability at least $1 - \delta$, outputs a distribution $Z \in S_{n,\epsilon}$ that satisfies $d_{TV}(X, Z) \leq 6\epsilon$.

Proof. The algorithm performs a tournament, by running $Choose - Hypothesis^X(H_1, H_2, \epsilon, \delta/(4N))$ for every pair $(H_i, H_j), i < j$, of distributions in $S_{n,\epsilon}$. Then it outputs any distribution $Y_* \in S_{n,\epsilon}$ that was never a loser (i.e., won or tied against all other distributions in the cover). If no such distribution exists in $S_{n,\epsilon}$ then the algorithm says "failure", and outputs an arbitrary distribution from $S_{n,\epsilon}$.

Since $S_{n,\epsilon}$ is an ϵ-cover of S_n, there exists some $Y \in S_{n,\epsilon}$ such that $d_{TV}(X, Y) \leq \epsilon$. We first argue that with high probability this distribution Y never loses a competition against any other $Y' \in S_{n,\epsilon}$ (so the algorithm does not output "failure"). Consider any $Y' \in S_{n,\epsilon}$. If $d_{TV}(X, Y') > 4\epsilon$, by Claim 3(ii) the probability that Y loses to Y' is at most $2e^{-m\epsilon'^2/2} \leq \delta/2N$. On the other hand, if $d_{TV}(X, Y') \leq 4\epsilon$, the triangle inequality gives that $d_{TV}(Y, Y') \leq 5\epsilon$ and thus Y draws against Y'. A union bound over all $N - 1$ distributions in $S_{n,\epsilon} - \{Y\}$ shows that with probability at least $1 - \delta/2$, the distribution Y never loses a competition.

We next argue that with probability at least $1 - \delta/2$, every distribution $Y' \in S_{n,\epsilon}$ that never loses must be close to X. Fix a distribution Y' such that $d_{TV}(X, Y') > 6\epsilon$. Claim 3(i) implies that Y' loses to Y with probability at least $1 - 2e^{-m\epsilon'^2/2} \geq 1 - \delta/(2N)$. A union bound gives that with probability at least $1 - \delta/2$, every distribution Y' that has $d_{TV}(X, Y') > 6\epsilon$ loses some competition. Thus, with overall probability at least $1 - \delta$, the tournament does not output "failure" and outputs some distribution Y^* such that $d_{TV}(X, Y^*) \leq 6\epsilon$. This proves the lemma.

Chapter 5

Differential Privacy on Learning Algorithm

In this section we examine the differential privacy on $Learning - PBD$ X algorithm. We first give an overview of the algorithm so as to remember and better understand its process. Then we will examine if the algorithm is differential private.

The algorithm tries to predict a PBD X by gaining access to a sample of the distribution. By theorem 3.3 we know that X will be ϵ-close to a sparse form PBD Y or ϵ-close to a (n, k)-Binomial form. To compute the approximation distribution, the algorithm performs 3 stages. In the first stage it use Birge algorithm, as a subroutine, and outputs a hypothesis distribution H_S with the following guarantee: if X is ϵ-close to some sparse form PBD Y then with probability $1 - \delta$ the hypothesis distribution H_S is ϵ-close the X, $d_{TV}(X, H_S) \leq c_1\epsilon$ for some constant $c_1 \geq 1$. On stage 2 the algorithm outputs a hypothesis distribution H_P with the following guarantee: if X is not ϵ-close to any sparse form PBD in the cover $S_{\epsilon'}$ the algorithm will output two estimation parameters μ, σ such that $H_P = TP(\mu, \sigma^2)$ (where $TP(\mu, \sigma^2)$ is the translated Poisson distribution) and with probability at least $1 - \delta$ we have

$d_{TV}(X, H_S) \leq c_2\epsilon$ for some constant $c_2 \geq 1$. Recall that from Theorem 3.3 X will be ϵ-close on a sparse PBD Y or on a (n, k)-heavy binomial form, thus we expect that X will be ϵ-close to H_S or H_P. In stage 3 the algorithm will decide in which of the two hypothesis distributions the PBD X is closer and with high probability will output the specific distribution. At the end the algorithm perform's a Tournament: For every pair $(H_i, H_j), i < j$ of hypothesis distributions in the cover S_ϵ it calculates the distance from the initial PBD X and with high probability outputs the "closer" one. This decision is made by running the subroutine $Choose - Hypothesis^X$. $Choose - Hypothesis^X$ compares the hypothesis distributions H_P and H_S with a sample of X and decides which one is closer. With high probability the subroutine will export the closest distribution.

As we mentioned above in this section we try to prove if and when the $Learn-PBD$ (we will refer to algorithm as $(Learn(X))$ hereafter) algorithm is differential private. We will try to prove that with high probability, a small change (by at most one entry) on algorithm's input dataset will affect negligible its output. More specific, assume two PBDs $X = (p_1, p_2, ..., p_n)$ and $X' = (p_1^{'}, p_2, ..., p_n)$ differ in **only one** entry (i.e. $p_1 \neq p_1^{'}$). Then with high probability the algorithm's output should remain ϵ-close. As stated above the algorithm computes its output in 3 steps. Thus, to conclude its differential privacy we have to ensure that in each step its output remain ϵ-close with high probability for both datasets X and X'. To this end we will first try to prove that with high probability the $Learn - Sparse^X$ algorithm outputs the same hypothesis distribution H_S for both X, X' datasets. The same idea applies for the subroutines $Learn - Poisson^X$ and $Choose - Hypothesis^X$. Finally we will also ensure that the Tournament performed at the end maintains the privacy. The results are quite interesting. More specific we will see that if the PBD X is close to a (n, k)-heavy binomial form then the algorithm becomes differential private (the hypothesis distribution H_P remains the same for both X, X'). However on case were X is close to a k-Sparse form the privacy depends on m (where m the number of $p_i^{'} \neq \{0, 1\}$ see Theorem 3.1). In our results we assume that the difference between the two PBDs X, X' datasets is from 0 to 1 (i.e. $p_1 = 0 \quad \text{and} \quad p_1^{'} = 1$ which represents the worst case, the highest mean difference).

We first try to define the difference between the two output hypothesis distributions regarding the $Learn - Sparse^X$ algorithm when X is close to a k-Sparse form. More specific assume two PBD X and X' as above. The subroutine $Learn - Sparse^X$ will output two hypothesis distributions H_S and $H_S^{'}$ ϵ-close to X and X'. We must calculate the variation distance of the latter. Observe that if we calculate the $d_{TV}(X, X')$ then by triangle inequality we will obtain an upper bound:

$$d_{TV}(H_S, H_S^{'}) \leq d_{TV}(H_S, X) + d_{TV}(X, X') + d_{TV}(X', H_S^{'}) \leq 2c\cdot\epsilon + d_{TV}(X, X').$$

We first prove that $d_{TV}(X, X') = P(X = t)$, where $t = \lfloor \mu \rfloor$ and μ the mean value of X. However, as an upper bound for $P(X = t)$ is quite difficult we will try to bound a relative probability $P(Y = np)$ where Y is a binomial distribution $Bin(n, p)$ and np its mean.

Theorem 5.1. *Assume $X = (p_1, p_2, ..., p_n)$ and $X' = (p_1^{'}, p_2, ..., p_n)$ two PBD differ in only one entry by 1 (i.e. $p_1 = 0 \quad \text{and} \quad p_1^{'} = 1$). Then $d_{TV}(X, X') = P(X = t)$, where $t = \lfloor \mu \rfloor$ and μ the mean value of X.*

Proof. We first observe that if $W_1 = \{0, 1, 2, ..., n\}$ is the support of X then $W_2 = \{1, 2, ..., n + 1\}$ will be the support of X' because of the change from 0 to 1 of the p_1 probability. By the definition of PBD mass function we observe that $P(X = k) = P(X' = k + 1)$, for $k = 0, .., n$. Indeed, $P(X = k) = \sum_{A \in F_k} \prod_{i \in A} p_i \prod_{i \in A^c} (1 - p_i)$, where F_k is the set of all subsets of k integers that can be selected from $\{1, 2, 3, ..., n\}$. Thus, if $p_1 = 0$ or $p_1 = 1$ these two probabilities do not contribute in the calculation of $P(X = k)$ value (they will give zero mass). Hence, considering that all others probabilities are equal $p_i = p_i^{'}, i \neq 1$ their mass functions will be exactly the same for each value in their respective domains W_1, W_2. Thus,

$$P(X = k) = P(X' = k + 1), \quad \text{for} \quad k = 0, .., n. \tag{5.1}$$

Finally, also by PBD definition it is easy to observe that

$$P(X = k-1) \leq P(X = k) \quad for \quad k = 1, ..., \lfloor \mu \rfloor \tag{5.2}$$

and

$$P(X = k) \geq P(X = k+1) \quad \text{for} \quad k = \lfloor \mu \rfloor, ..., n \tag{5.3}$$

where μ the mean value of X.

Let $W = W_1 \cup W_2$, $t = \lfloor \mu \rfloor$, where $\mu = \sum_i p_i$ the mean value of X (thus, $\mu + 1$ the mean value of X'), we prove that:

$$d_{TV}(X, X') = \frac{1}{2} \sum_{w \in W} (|P(X = w) - P(X' = w)|)$$

$$= \frac{1}{2} \sum_{w \in W} (|P(X' = w+1) - P(X' = w)|) \quad \text{from (5.1)}$$

$$= \frac{1}{2}(|P(X' = 1) - P(X' = 0)| + ... + |P(X' = t+1) - P(X' = t)|$$

$$+|P(X' = t+2) - P(X' = t+1)| + ... + |P(X = n+1) - P(X' = n+1)|)$$

$$= \frac{1}{2}(P(X' = 1) + P(X' = 2) - P(X' = 1) + ... + P(X' = t+1) - P(X' = t)$$

$$+P(X' = t+1) - P(X' = t+2) + ... + P(X' = n+1)) \quad \text{from (5.2),(5.3)}$$

$$= P(X' = t+1) = P(X = t).$$

As we discussed we will try to bound a relative probability $P(Y = np)$.

Theorem 5.2. *Assume Y a binomial distribution $Bin(n, p)$ and np its mean. We will show that $P(Y = np) \leq \frac{e}{2\pi} \cdot \frac{1}{\sqrt{n} \cdot \sqrt{p(1-p)}}$*

Proof. We will first give a lemma which will help us to prove the theorem.

Lemma 5.3. *Stirling's Approximation*
For all positive integers n the following inequality holds:

$$\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \leq n! \leq e n^{n+\frac{1}{2}} e^{-n}$$

We now begin the proof of the theorem:

$$P(y = np) = \left(\begin{array}{c} n \\ np \end{array} \right) \cdot p^{np} \cdot (1-p)^{n-np} = \frac{n!}{(np)!(n-np)!} \cdot p^{np} \cdot (1-p)^{n-np}$$

$$\leq \frac{en^{n+\frac{1}{2}}e^{-n} \cdot p^{np} \cdot (1-p)^{n-np}}{\sqrt{2\pi}(np)^{np+\frac{1}{2}}e^{-np}\sqrt{2\pi}(n-np)^{(n-np)+\frac{1}{2}}e^{-(n-np)}}$$

$$= \frac{\frac{en^n\sqrt{n}\cdot p^{np}\cdot(1-p)^{n-np}}{e^n}}{\frac{2\pi(np)^{np}\sqrt{np}(n-np)^{n-np}\sqrt{n-np}}{e^{np}e^{n-np}}}$$

$$= \frac{en^n\sqrt{n} \cdot p^{np} \cdot (1-p)^{n-np}}{2\pi n^{np} \cdot p^{np} \cdot n^{n-np} \cdot (1-p)^{n-np}\sqrt{n^2p - n^2p^2}}$$

$$= \frac{e}{2\pi} \cdot \frac{\sqrt{n}}{n\sqrt{p(1-p)}}$$

$$= \frac{e}{2\pi} \cdot \frac{1}{\sqrt{n}\sqrt{p(1-p)}}$$

Thus as we discussed the bound depends on distributions cardinality n or equivalent by its variance as $np(1-p) = \sigma^2$.

As a next step we will prove that with high probability the hypothesis distributions output of the algorithm $Learn - Poisson^X$ between X and X' are ϵ-close.

Theorem 5.4. *Let X, X' be two PBD of order n, differ in only one entry. We can assume that the two distributions are of the form $X = (p_1, ..., p_n)$, $X' = (p_1^{'}..., p_n)$, and $p_1 = 0, p_1^{'} = 1$. We define $\mu = \mathbb{E}[X], \mu' = \mathbb{E}[X']$ and $\sigma^2 = Var[X], \sigma^{'2} = Var[X']$. Let $\hat{\mu}, \hat{\mu'}, \hat{\sigma}^2, \hat{\sigma}^{'2}$ be the estimated means and variances of X, X' respectively produced by the algorithm. We will prove that for all $n, \epsilon, \delta > 0$,*

$$d_{TV}(TP(\hat{\mu}, \hat{\sigma}^2), TP(\hat{\mu'}, \hat{\sigma}^{'2})) \leq O(\epsilon)$$

With probability at least $(1-\delta)^2$.

Proof. As it has been described in subsection 4.1.2 the algorithm $Learn - Poisson^X$ will output a translated Poisson distribution $TP(\hat{\mu}, \hat{\sigma}^2)$ with parameters $\hat{\mu}, \hat{\sigma}^2$. Thus the algorithm will output two translated Poisson distributions $TP(\hat{\mu}, \hat{\sigma}^2), TP(\hat{\mu}', \hat{\sigma}^{'2})$, one for each X and X' respectively. For those parameters $\mu, \mu', \sigma^2, \sigma^{'2}$ the properties: (4.1) and (4.2) from section 4.1.2 holds. Thus with probability at least $1 - \delta$:

$$\sigma^2, \sigma^{'2} = \Omega(1/\epsilon^{'2}) \geq \theta^2 \tag{5.4}$$

$$|\mu - \hat{\mu}| \leq \epsilon' \sigma, |\sigma^2 - \hat{\sigma}^2| \leq \epsilon' \sigma^2 \tag{5.5}$$

$$|\mu' - \hat{\mu}'| \leq \epsilon' \sigma', |\sigma^{'2} - \hat{\sigma}^{'2}| \leq \epsilon' \sigma^{'2} \tag{5.6}$$

We first bound the $|\mu - \mu'|$ and $|\sigma^2 - \sigma^{'2}|$. We prove that the distance between the two mean values μ, μ' is:

$$|\mu - \mu'| = \left| \sum_{i=1}^{n} p_i - \sum_{i=2}^{n} p_i + p_1^{'} \right| = \left| (p_1 - p_1^{'}) \right| = 1.$$

When the distance between the two variances is

$$|\sigma - \sigma'| = \left| \sum_{i=1}^{n} (1 - p_i) p_i - \left(\sum_{i=1}^{n} (1 - p_i^{'}) p_i^{'} \right) \right| =$$

$$= \left| \sum_{i=2}^{n} (1 - p_i) p_i - \left(\sum_{i=2}^{n} (1 - p_i^{'}) p_i^{'} \right) \right|$$

$$= \left| \sum_{i=2}^{n} (1 - p_i) p_i - \left(\sum_{i=2}^{n} (1 - p_i) p_i \right) \right| = 0.$$

Thus

$$|\hat{\mu} - \hat{\mu}'| \leq |\hat{\mu} - \mu| + |\mu - \mu'| + |\mu' - \hat{\mu}'| \leq \epsilon' \cdot \sigma + \epsilon' \cdot \sigma' + 1$$

$$= \epsilon' \cdot (\sigma + \sigma') + 1 = 2\epsilon' \cdot \sigma + 1.$$

and

$$|\hat{\sigma}^2 - \hat{\sigma}^{'2}| \leq |\hat{\sigma}^2 - \sigma^2| + |\sigma^2 - \sigma^{'2}| + |\sigma^{'2} - \hat{\sigma}^{'2}| \leq \epsilon \cdot \sigma^2 + \epsilon \cdot \sigma^{'2}$$

$$= 2\epsilon' \cdot \sigma^2$$

With probability at least $(1-\delta)^2$.

We then bound the total variation distance between the two translated Poisson distributions:

$$d_{TV}(TP(\hat{\mu}, \hat{\sigma}^2), TP(\hat{\mu'}, \hat{\sigma}'^2)) \leq \frac{|\hat{\mu} - \hat{\mu'}|}{min(\hat{\sigma}, \hat{\sigma}')} + \frac{|\hat{\sigma}^2 - \hat{\sigma}'^2| + 1}{min(\hat{\sigma}^2, \hat{\sigma}'^2)}$$

$$\leq \frac{2\epsilon'\hat{\sigma} + 1}{\sigma} + \frac{2\epsilon'\hat{\sigma}^2 + 5/4}{\sigma^2}$$

$$\leq \frac{2\epsilon'\hat{\sigma} + 1}{\sigma/\sqrt{1-\epsilon}} + \frac{2\epsilon'\hat{\sigma}^2 + 5/4}{\sigma^2/(1-\epsilon)}$$

$$= O(\epsilon) + O(\epsilon^2)$$

$$= O(\epsilon).$$

Hence the output of the $Learn - Poisson^X$ algorithm for both X and X' is ϵ-close.

As a last step it remains to show that the Tournament with high probability will output the "right" hypothesis distributions. As we saw the subroutines $Learn-Sparse^X$ and $Learn-Poisson^X$ will output two hypothesis distributions H_S, H_P close to the initial distribution X. The same will apply for the PBD X', obtaining $H_S^{'}, H_P^{'}$. The algorithm in his last step will perform a competition between the two candidates hypothesis distributions (H_S, H_P for X and $H_S^{'}, H_P^{'}$ for X') and with hight probability $(1-2e^{-m\epsilon^2/2})$ will output the closest one. Thus, for both X, X' the $Choose-Hypothesis^X$ subroutine will output the closest distributions (in respect to X and X') with probability at least $(1-2e^{-m\epsilon^2/2})^2$.

As a last step we run the $Choose-Hypothesis^X$ algorithm for every pairs $(H_i, H_j), i < j$ of distributions in S_ϵ. Then it output, with high probability (at least $1-\delta$) a distribution $Z \in S_\epsilon$ that was never a looser and with total variation at most 6ϵ i.e. $d_{TV}(X, Z) \leq 6\epsilon$. Thus, we must show that the

Tournament will keep "close" its output with high probability for both X and X'.

Lemma 5.5. *Let $X = (p_1, p_2, ..., p_n)$ and $X' = (p_1^{'}, p_2, ..., p_n)$ be two PBDs differ in only one entry by 1 (i.e. $p_1 = 0, p_1^{'} = 1$). Let also Z and Z' be the two outputs of the Tournament for X and X' respectively. Then with probability $(1-\delta)^2$ their variation distance will be $d_{TV}(Z, Z') \leq 12\epsilon + d$, where $d = d_{TV}(X, X')$.*

Proof. The proof is simply as it comes directly from the triangle inequality. With probability at least $1-\delta$ the Tournament (for X) will output a distribution Z such that $d_{TV}(X, Z) \leq 6\epsilon$. The same holds for X'. The Tournament will output a distribution Z' such that $d_{TV}(X', Z') \leq 6\epsilon$.

Then from the triangle inequality:

$$\begin{aligned} d_{TV}(Z, Z') &\leq d_{TV}(Z, X) + d_{TV}(X, X') + d_{TV}(X', Z') \\ &\leq 6\epsilon + d + 6\epsilon \\ &= 12\epsilon + d, \end{aligned}$$

with probability at least $(1-\delta)^2$.

As we show the Tournament will output with high probability two distributions Z, Z' which are $12\epsilon + d$ close. The variation distance d of X, X' depends on the nature of the PBD distributions X and X'. If X and X' are close to a k-Sparse form then their variation distance depends on their cardinality n. If X, X' are close to a (n, k)-Binomial form then their distance remains ϵ-close. The following section gives a brief description of our results.

Chapter 6

Conclusions \ Next Steps

As we show the algorithm Lean Poisson Binomial Distribution performs differential privacy with respect to the following conditions:

- If the PBD X is close to a (n, k)-Binomial Form then the algorithm is differential private
- On case where X is close to a k-Sparse form the property of differential privacy depends on the PBD cardinality

As next steps the following may be considered:

- Give a lower variation distance bound so as to show if the algorithm is optimal (by its construction) regarding its Differential Privacy property.
- Provide a better bound regarding $P(X = t)$ theorem's 5.1 output
- Add noise on Algorithm's subroutine $Learn - Sparse$ to maintain privacy for sparse cardinality

Bibliography

[1] K.Pearson. *Contribution to the Mathematical Theory of Evolution.* Philosophical Transaction of the Royal Society in London 1894

[2] D. Titterington, A. Smith, U. Markov *Statistical Analysis of Finite Mixture Distributions* Wiley 1985

[3] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. Schapire, L. Sellie. *On the Learnability of Discrete Distributions.* ACM Symposium on Theory of Computing, 1994

[4] L. Birge. *Estimation of unimodal densities without smoothness assumptions.* Annals of Statistics, 25(3):970-981, 1997.

[5] C. Daskalakis, G. Kamath *Faster and Sample Near-Optimal Algorithms for Proper Learning Mixtures of Gaussians.* Annual Conference on Learning Theory, 2014

[6] C. Daskalakis, I. Diakonikolas, R. Servedio *Learning Poisson Binomial Distributions.* ACM Symposium on Theory of Computing, 2012

[7] C. Daskalakis, C. Papadimitriou *Sparse Covers for Sums of Indicators.* Probability Theory and Related Fields, 2014

[8] C. Daskalakis, I. Diakonikolas, R. O'Donnell, R. Servedio, L. Tan *Learning Sums of Independent Integer Random Variables.* IEEE Symposium on Foundations of Computer Science, 2013

[9] G.A. Carpenter, S. Grossberg *The ART of adaptive pattern recognition by a self-organizing neural network.* 1988

[10] Sanjeev Kulkarni, Gilbert Harman *An Elementary Introduction to Statistical Learning Theory.* 2011

[11] Andrew R. Webb, Keith D. Copsey *Statistical Pattern Recognition.* 2011

[12] P. Samarati and L. Sweeney. *Generalizing data to provide anonymity when disclosing information.* page 188, 1998. cited By (since 1996)

[13] L. Sweeney *k-anonymity: A model for protecting privacy.* International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems, 10(5):557-570, 2002. cited By (since 1996)

[14] C. Dwork *Differential privacy.* In ICALP'06: Proceedings of the 33rd international conference on Automata, Languages and Programming, pages 1-12, Berlin, Heidelberg, 2006.

[15] A. Narayanan and V. Shmatikov *Robust de-anonymization of large sparse datasets.* In Proceedings of the 2008 IEEE Symposium on Security and Privacy, SP '08, pages 111-125, Washington, DC, USA, 2008. IEEE Computer Society.

[16] M. Srivatsa and M. Hicks *Deanonymizing mobility traces: using social network as a side-channel.* In Proceedings of the 2012 ACM conference on Computer and communications security, CCS '12, pages 628-637, New York, NY, USA, 2012

[17] S. Le Blond, C. Zhang, A. Legout, K. Ross, and W. Dabbous. *I know where you are and what you are sharing: exploiting P2P communications to invade users' privacy.* In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, IMC '11, pages 45-60, New York, NY, USA, 2011

[18] 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, New York, NY, 2006, pp. 265-284

[19] F. McSherry and K. Talwar, *Mechanism design via differential privacy.* in Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, Providence, RI, 2007, pp. 94-103

[20] P. Samarati and L. Sweeney *Generalizing data to provide anonymity when disclosing information.* page 188, 1998. cited By (since 1996)

[21] L. Sweeney *k-anonymity: A model for protecting privacy.* International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems, 10(5):557-570, 2002. cited By (since 1996)

[22] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam *L-diversity: Privacy beyond k-anonymity.* ACM Trans. Knowl. Discov. Data, 1(1), Mar. 2007

[23] N. Li, T. Li, and S. Venkatasubramanian *t-closeness: Privacy beyond k-anonymity and l-diversity.* In Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on, pages 106-115, april 2007

[24] N. Li, T. Li, and S. Venkatasubramanian *Closeness: A new privacy measure for data publishing.* Knowledge and Data Engineering, IEEE Transactions on, 22(7):943-956, july 2010

[25] M. E. Nergiz, M. Atzori, and C. Clifton *Hiding the presence of individuals from shared databases.* In Proceedings of the 2007 ACM SIGMOD international conference on Management of data, SIGMOD '07, pages 665-676, New York, NY, USA, 2007

[26] Sweeney L. *Replacing Personally Identifiable Information in Medical Records, the Scrub System.* Journal of the American Medical Informatics Association, 1996

[27] Sweeney L. *Guaranteeing Anonymity while Sharing Data, the Datafly System.* Journal of the American Medical Informatics Association, 1997

[28] Sweeney L. *Privacy Technologies for Homeland Security.* Testimony before the Privacy and Integrity Advisory Committee of the Department of Homeland Security, Boston, MA, June 15, 2005

[29] A. Kalai, A. Moitra *G. Valiant Efficiently Learning Mixtures of Two Gaussians* ACM Symposium on Theory of Computing, 2010

[30] L. Valiant *A theory of the learnable.* Communications of ACM, 1984

[31] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun *Learning mixtures of structured distributions over discrete domains.* In SODA, pages 1380-1394, 2013

[32] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun *Efficient density estimation via piecewise polynomial approximation.* In STOC, pages 604-613, 2014.

[33] S. Chan, I. Diakonikolas, R. Servedio, and X. Sun *Near-optimal density estimation in near-linear time using variable-width histograms.* In NIPS, pages 1844-1852, 2014

[34] R. Chen, B. C. Fung, B. C. Desai, and N. M. Sossou *Differentially private transit data publication: a case study on the Montreal transportation system.* In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '12, pages 213-221, New York, NY, USA, 2012

[35] M. Gotz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke. *Publishing search logs: A comparative study of privacy guarantees.* IEEE Transactions on Knowledge and Data Engineering, 24(3):520, 2012

[36] F. McSherry and I. Mironov *Differentially private recommender systems: building privacy into the Netflix Prize Contenders.* Proceedings of the

15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) pages 627-636, 2009

[37] H. Chernoff *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations.* Ann. Math. Statist., 23:493-507, 1952.

[38] W. Hoeffding *Probability inequalities for sums of bounded random variables.* Journal of the American Statistical Association, 58:13-30, 1963

[39] D. Dubhashi and A. Panconesi *Concentration of measure for the analysis of randomized algorithms.* Cambridge University Press, Cambridge, 2009

[40] S.X. Chen and J.S. Liu *Statistical applications of the Poisson-Binomial and Conditional Bernoulli Distributions.* Statistica Sinica, 7:875-892, 1997

[41] C. Daskalakis and C. Papadimitriou *On Oblivious PTAS's for Nash Equilibrium.* STOC 2009, pp. 75-84. Full version available as ArXiV report, 2011

[42] C. Daskalakis *An Efficient PTAS for Two-Strategy Anonymous Games.* WINE 2008, pp. 186-197. Full version available as ArXiV report, 2008

[43] A. Rollin *Translated Poisson Approximation Using Exchangeable Pair Couplings.* Annals of Applied Probability, 17(5/6):1596-1614, 2007

[44] J. Keilson and H. Gerber. *Some Results for Discrete Unimodality.* J. American Statistical Association, 66(334):386-389, 1971

[45] L. Devroye and G. Lugosi *A universally acceptable smoothing factor for kernel density estimation.* Annals of Statistics, 24:2499-2512, 1996

[46] L. Devroye and G. Lugosi *Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes.* Annals of Statistics, 25:2626-2637, 1996

[47] L. Devroye and G. Lugosi *Combinatorial methods in density estimation.* Springer Series in Statistics, Springer, 2001

[48] Y. G. Yatracos *Rates of convergence of minimum distance estimators and Kolmogorov's entropy.* Annals of Statistics, 13:768-774, 1985

[49] Andrew D. Barbour and Torgny Lindvall *Translated Poisson Approximation for Markov Chains.* Journal of Theoretical Probability, 19(3):609-630, 2006

[50] Werner Ehm *Binomial Approximation to the Poisson Binomial Distribution.* Statistics and Probability Letters, 11:7-16, 1991

[51] *S. Kotz N.L. Johnson, A.W. Kemp* Univariate discrete distributions. John Wiley Sons, Inc., New York, NY, USA, 2005