Algebraic and Analytic Approaches for Parameter Learning in Mixture Models

Akshay Krishnamurthy AKSHAY@CS.UMASS.EDU
Microsoft Research, New York City, NY 10011

Arya Mazumdar ARYA@CS.UMASS.EDU
University of Massachusetts Amherst, MA 01003, USA

Andrew McGregor MCGREGOR@CS.UMASS.EDU
University of Massachusetts Amherst, MA 01003, USA

Soumyabrata Pal SPAL@CS.UMASS.EDU
University of Massachusetts Amherst, MA 01003, USA

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

We present two different approaches for parameter learning in several mixture models in one dimension. Our first approach uses complex-analytic methods and applies to Gaussian mixtures with shared variance, binomial mixtures with shared success probability, and Poisson mixtures, among others. An example result is that $\exp(O(N^{1/3}))$ samples suffice to exactly learn a mixture of $k < N$ Poisson distributions, each with integral rate parameters bounded by N. Our second approach uses algebraic and combinatorial tools and applies to binomial mixtures with shared trial parameter N and differing success parameters, as well as to mixtures of geometric distributions. Again, as an example, for binomial mixtures with k components and success parameters discretized to resolution ϵ, $O(k^2(N/\epsilon)^{8/\sqrt{\epsilon}})$ samples suffice to exactly recover the parameters. For some of these distributions, our results represent the first guarantees for parameter estimation.

Keywords: Parameter learning, mixture model, complex analysis, method of moments.

1. Introduction

Mixture modeling is a powerful method in the statistical toolkit, with widespread use across the sciences (Titterington et al., 1985). Starting with the seminal work of Dasgupta (1999), computational and statistical aspects of learning mixture models have been the subject of intense investigation in the theoretical computer science and statistics communities (Achlioptas and McSherry, 2005; Kalai et al., 2010; Belkin and Sinha, 2010; Arora and Kannan, 2001; Moitra and Valiant, 2010; Feldman et al., 2008; Chan et al., 2014; Acharya et al., 2017; Hopkins and Li, 2018; Diakonikolas et al., 2018; Kothari et al., 2018; Hardt and Price, 2015).

In this literature, there are two flavors of result: (1) *parameter estimation*, where the goal is to identify the mixing weights and the parameters of each component from samples, and (2) *density estimation* or PAC-learning, where the goal is simply to find a distribution that is close in some distance (e.g., TV distance) to the data-generating mechanism. Density estimation can be further subdivided into *proper* and *improper learning* approaches depending on whether the algorithm outputs a distribution from the given mixture family or not. These three guarantees are quite different. Apart from Gaussian mixtures, where all types of results exist, prior work for other mixture families

largely focuses on density estimation, and very little is known for parameter estimation outside of Gaussian mixture models. In this paper, we focus on parameter estimation and provide two new approaches, both of which apply to several mixture families.

Our first approach is analytic in nature and yields new sample complexity guarantees for univariate mixture models including Gaussian, Binomial, and Poisson. Our key technical insight is that we can relate the total variation between two candidate mixtures to a certain Littlewood polynomial, and then use complex analytic techniques to establish separation in TV-distance. With this separation result, we can use density estimation techniques (specifically proper learning techniques) to find a candidate mixture that is close in TV-distance to the data generating mechanism. The results we obtain via this approach are labeled as "analytic" in Table 1. This approach has recently led to important advances in the trace reconstruction and population recovery problems; see work by De et al. (2017a), Nazarov and Peres (2017), and De et al. (2017b).

Our second approach is based on the *method of moments*, a popular approach for learning Gaussian mixtures, and is more algebraic. Roughly, these algorithms are based on expressing moments of the mixture model as polynomials of the component parameters, and then solving a polynomial system using estimated moments. This approach has been studied in some generality by Belkin and Sinha (2010) who show that it can succeed for a large class of mixture models. However, as their method uses non-constructive arguments from algebraic geometry it cannot be used to bound how many moments are required, which is essential in determining the sample complexity; see a discussion in (Moitra, 2018, Section 7.6). In contrast, our approach does yield bounds on how many moments suffice and can be seen as a quantified version of the results in Belkin and Sinha (2010). The results we obtain via this approach are labeled as "algebraic" in Table 1.

The literature on mixture models is quite large, and we have just referred to a sample of most relevant papers here. A bigger overview on learning distributions can be found in the recent monographs such as Moitra (2018); Diakonikolas (2016).

1.1. Overview of results

As mentioned, an overview of our sample complexity results are displayed in Table 1, where in all cases we consider a uniform mixture of k distributions. Our guarantees are for *exact* parameter estimation, under the assumption that the mixture parameters are discretized to a particular resolution, given in the third column of the table. Theorem statements are given in the sequel.

At first glance the guarantees seem weak, since they all involve exponential dependence in problem parameters. However, except for the Gaussian case, these results are the first guarantees for parameter estimation for these distributions. All prior results we are aware of consider density estimation (Chan et al., 2013; Feldman et al., 2008).

For the mixtures of discrete distributions, such as binomial and negative binomial with shared trial parameter, or Poisson/geometric/chi-squared mixtures with certain discretizations, it seems like the dependence of sample complexity on the number of components k is polynomial (see Table 1). Note that for these examples $k \leq N$, the upper bounds on parameter values. Therefore the actual dependence on k can still be interpreted as exponential. The results are especially interesting when k is large and possibly growing with N.

For Gaussian mixtures, the most interesting aspect of our bound is the polynomial dependence on the number of components k (first row of Table 1). In our setting and taking $\sigma = 1$, the result of Moitra and Valiant (2010) is applicable, and it yields $\epsilon^{-O(k)}$ sample complexity, which is incom-

Distribution	Pdf/Pmf $f(x;\theta)$	Discretization	Sample Complexity	Approach
Gaussian	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{\sigma^2}}$	$\mu_i \in \epsilon\mathbb{Z}$	$k^3\exp(O((\sigma/\epsilon)^{2/3}))$	Analytic
Binomial	$\binom{n}{x}p^x(1-p)^{n-x}$	$n_i \in \{1,2,\ldots,N\}$	$\exp(O(((N/p)^{1/3})))$	Analytic [2]
		$p_i \in \{0,\epsilon,\ldots,1\}$	$O(k^2(n/\epsilon)^{8/\sqrt{\epsilon}})$	Algebraic
Poisson	$\frac{\lambda^x e^{-\lambda}}{x!}$	$\lambda_i \in \{0,1,\ldots,N\}$	$\exp(O(N^{1/3}))$	Analytic
Geometric	$(1-p)^x p$	$1/p_i \in \{1,\ldots,N\}$	$O(k^2(\sqrt{N})^{8\sqrt{N}})$	Algebraic
		$p_i \in \{0,\epsilon,\ldots,1\}$	$O(\frac{k^2}{\epsilon^{8/\sqrt{\epsilon}+2}}\log\frac{1}{\epsilon})$	Algebraic
χ^2	$\frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}$	$n_i \in \{0,1,\ldots,N\}$	$\exp(O(N^{1/3}))$	Analytic
Negative Binomial	$\binom{x+r-1}{x}(1-p)^r p^x$	$r_i \in \{1,2,\ldots,N\}$	$\exp(O((N/p)^{1/3}))$	Analytic

Table 1: Overview of our results. Results are given for uniform mixtures of k different components but some can be extended to non-uniform mixtures. Note that for rows 2, 4, 7, and 8, k does not appear. This is because $k \leq N$ and other terms dominate.

parable to our $k^3\exp(O(\epsilon^{-2/3}))$ bound. Note that our result avoids an exponential dependence in k, trading this off for an exponential dependence on the discretization/accuracy parameter ϵ.[1] Other results for Gaussian mixtures either 1) consider density estimation (Daskalakis and Kamath, 2014; Feldman et al., 2008), which is qualitatively quite different from parameter estimation, 2) treat k as constant (Hardt and Price, 2015; Kalai et al., 2010), or 3) focus on the high dimensional setting and require separation assumptions (see for example Diakonikolas et al. (2017) and Moitra (2018)).

As such, our results reflect a new sample complexity tradeoff for parameter estimation in Gaussian mixtures.

As another note, using ideas from (Nazarov and Peres, 2017; De et al., 2017a), one can show that the analytic result for Binomial mixtures is optimal. This raises the question of whether the other results are also optimal or is learning a Binomial mixture intrinsically harder than learning, e.g., a Poisson or Gaussian mixture?

As a final remark, our assumption that parameters are discretized is related to separation conditions that appear in the literature on learning Gaussian mixtures. However, our approach does not seem to yield guarantees when the parameters do not exactly fall into the discretization. We hope to resolve this shortcoming in future work.

1. Due to our discretization structure, our results do not contradict the lower bounds of Moitra and Valiant (2010); Hardt and Price (2015).

1.2. Our techniques

To establish these results, we take two loosely related approaches. In our analytic approach, the key structural result is to lower bound the total variation distance between two mixtures $\mathcal{M}, \mathcal{M}'$ by a certain Littlewood polynomial. For each distribution type, if the parameter is θ, we find a function $G_t : \mathbb{R} \to \mathbb{C}$ such that

$$\mathbb{E}[G_t(X)] = \exp(it\theta).$$

(For Gaussians, G_t is essentially the characteristic function). Such functions can be used to obtain Littlewood polynomials from the difference in expectation for two different mixtures, for example if the parameters θ are integral and the mixture weights are uniform. Applying complex analytic results on Littlewood polynomials, this characterization yields a lower bound on the total variation distance between mixtures, at which point we may use density estimation techniques for parameter learning. Specifically we use the minimum distance estimator (see, Devroye and Lugosi (2012, Sec. 6.8)), which is based on the idea of Scheffe sets. Scheffe sets are building blocks of the Scheffe estimator, commonly used in density estimation, e.g. Suresh et al. (2014).

Our algebraic approach is based on the more classical method of moments. Our key innovation here is a combinatorial argument to bound the number of moments that we need to estimate in order to exactly identify the correct mixture parameters. In more detail, when the parameters belong to a discrete set, we show that the moments reveal various statistics about the multi-set of parameters in the mixture. Then, we adapt and extend classical combinatorics results on sequence reconstruction to argue that two distinct multi-sets must disagree on a low-order moment. These combinatorial results are related to the Prouhet-Tarry-Escott problem (see, e.g., Borwein (2002)) which also has connections to Littlewood polynomials. To wrap up we use standard concentration arguments to estimate all the necessary moments, which yields the sample complexity guarantees.

We note that the complex analytic technique provides non-trivial result only for those mixtures for which an appropriate function G_t exists. On the other hand, the algebraic approach works for all mixtures whose ℓ^{th} moment can be described as a polynomial of degree exactly ℓ in its unknown parameters. In Belkin and Sinha (2010), it was shown that most distributions have this later property. In general, where both methods can be applied, the complex analytic techniques typically provide tighter sample complexity bounds than the algebraic ones.

2. Learning Mixtures via Characteristic Functions

In this section, we show how analysis of the characteristic function can yield sample complexity guarantees for learning mixtures. At a high level, the recipe we adopt is the following.

1. First, we show that, in a general sense, the total variation distance between two separated mixtures is lower bounded by the L_∞ norm of their characteristic functions.

2. Next, we use complex analytic methods and specialized arguments for each particular distribution to lower bound the latter norm.

2. We obtained this result as a byproduct of sparse trace reconstruction (Krishnamurthy et al., 2019). In fact, the present work was motivated by the observation that the technique we were using there is much more general.

3. Finally, we use the minimum distance estimator (Devroye and Lugosi, 2012) to find a mixture that is close in total variation to the data generating distribution. Using uniform convergence arguments this yields exact parameter learning.

The two main results we prove in this section are listed below.

Theorem 1 (Learning Gaussian mixtures) *Let $\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k}\mathcal{N}(\mu_i, \sigma^2)$ be a uniform mixture of k univariate Gaussians, with known shared covariance σ^2 and with distinct means $\mu_i \in \epsilon\mathbb{Z}$. Then there exists an algorithm that requires $k^3\exp(O((\sigma/\epsilon)^{2/3}))$ samples from $\mathcal{M}$ and exactly identifies the parameters $\{\mu_i\}_{i=1}^{k}$ with high probability.*

Theorem 2 (Learning Poisson mixtures) *Let $\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k} Poi(\lambda_i)$ where $\lambda_i \in \{0, 1, \ldots, N\}$ for each i are distinct. Then there exists an algorithm that that requires $\exp(O(N^{1/3}))$ samples from $\mathcal{M}$ to exactly identify the parameters $\{\lambda_i\}_{i=1}^{k}$ with high probability.*

There are some technical differences in deriving the results for Gaussian vs Poisson mixtures. Namely, because of finite choice of parameters we can take a union bound over the all possible incorrect mixtures for the latter case, which is not possible for Gaussian. For Gaussian mixtures we instead use an approach based on VC dimension. The results for negative binomial mixtures and chi-squared mixtures (shown in Table 1) follow the same route as the Poisson mixture. As reported in Table 1, this approach also yields results for mixtures of binomial distributions that we obtained in a different context in our prior work (Krishnamurthy et al., 2019).

2.1. Total Variation and Characteristic Functions

Let $\{f_\theta\}_{\theta\in\Theta}$ denote a parameterized family of distributions over a sample space $\Omega \subset \mathbb{R}$, where f_θ denotes either a pdf or pmf, depending on the context. We call $\mathcal{M}$ a (finite) Θ-*mixture* if $\mathcal{M}$ has pdf/pmf $\sum_{\theta\in\mathcal{A}}\alpha_\theta f_\theta$ and $\mathcal{A} \subset \Theta, |\mathcal{A}| = k$. For a distribution with density f (we use distribution and density interchangeably in the sequel), define the *characteristic function* $C_f(t) \equiv \mathbb{E}_{X\sim f}[e^{itX}]$. For any two distribution f, f' defined over a sample space $\Omega \subseteq \mathbb{R}$ the variational distance (or the TV-distance) is defined to be $\|f - f'\|_{\text{TV}} \equiv \frac{1}{2}\int_\Omega \left|\frac{df'}{df} - 1\right| df$. For a function $G : \Omega \to \mathbb{C}$ define the L_∞ norm to be $\|G\|_\infty = \sup_{\omega\in\Omega}|G(\omega)|$ where $|\cdot|$ denotes the modulus.

As a first step, our aim is to show that the total variation distance between $\mathcal{M} = \sum_{\theta\in\mathcal{A}}\alpha_\theta f_\theta$ and any other mixture $\mathcal{M}'$ given by $\sum_{\theta\in\mathcal{B}}\beta_\theta f_\theta, \mathcal{B} \subset \Theta, |\mathcal{B}| = k$ is lower bounded. The following elementary lemma completes the first step of the outlined approach.

Lemma 3 *For any two distributions f, f' defined over the same sample space $\Omega \subseteq \mathbb{R}$, we have*

$$\left\|f - f'\right\|_{TV} \geq \frac{1}{2}\sup_{t\in\mathbb{R}}|C_f(t) - C_{f'}(t)|.$$

More generally, for any $G : \Omega \to \mathbb{C}$ and $\Omega' \subset \Omega$ we have

$$\left\|f - f'\right\|_{TV} \geq \left(2\sup_{x\in\Omega'}|G(x)|\right)^{-1}\Big(\left|\mathbb{E}_{X\sim f}G(X) - \mathbb{E}_{X\sim f'}G(X)\right| - \int_{x\in\Omega\setminus\Omega'}|G(x)|\cdot|df(x) - df'(x)|\Big).$$

Proof We prove the latter statement, which implies the former since for the function $G(x) = e^{itx}$ we have $\sup_x |G(x)| = 1$. We have

$$\begin{aligned}|\mathbb{E}_{X\sim f} G(X) - \mathbb{E}_{X\sim f'} G(X)| &\leq \int_{x\in\Omega} |G(x)| \cdot |df(x) - df'(x)| \\ &\leq 2 \sup_{x\in\Omega'} |G(x)| \cdot \left\|f - f'\right\|_{\text{TV}} + \int_{x\in\Omega\setminus\Omega'} |G(x)| \cdot |df(x) - df'(x)|.\end{aligned}$$

■

Equipped with the lower bound in Lemma 3, for each type of distribution, we set out to find a good function G to witness separation in total variation distance. As we will see shortly, for a parametric family f_θ, it will be convenient to find a family of functions G_t such that

$$\mathbb{E}_{X\sim f_\theta}[G_t(X)] = \exp(it\theta).$$

Of course, to apply Lemma 3, it will also be important to understand $\|G_t\|_\infty$. While such functions are specific to the parametric model in question, the remaining analysis will be unified. We derive such functions and collect the relevant properties in the following lemma. At a high level, the calculations are based on reverse engineering from the characteristic function, e.g., finding a choice $t'(t)$ such that $C_f(t') = \exp(it\theta)$.

Lemma 4 *Let $z = \exp(it)$ where $t \in [-\pi/L, \pi/L]$.*

- *Gaussian. If $X \sim \mathcal{N}(\mu, \sigma)$ and $G_t(x) = e^{itx}$ then*

$$\mathbb{E}[G_t(X)] = \exp(-\sigma^2 t^2/2) z^\mu \text{ and } \|G_t\|_\infty = 1\ .$$

- *Poisson. If $X \sim Poi(\lambda)$ and $G_t(x) = (1+it)^x$ then*

$$\mathbb{E}[G_t(X)] = z^\lambda \text{ and } |G_t(x)| \leq (1+t^2)^{x/2}\ .$$

- *Chi-Squared. If $X \sim \chi^2(\ell)$ and $G_t(x) = \exp(x/2 - xe^{-2it}/2)$ then*

$$\mathbb{E}[G_t(X)] = z^\ell \text{ and } |G_t(x)| \leq e^{cxt^2 + O(xt^4)}\ .$$

- *Negative Binomial. If $X \sim NB(r,p)$ and $G_t(x) = \left(1/p - (1/p-1)e^{-it}\right)^x$ then*

$$\mathbb{E}[G_t(X)] = z^r \text{ and } |G_t(x)| \leq e^{-cx\frac{(1-p)t^2}{p^2}}\ .$$

Proof Here we give the argument for Poisson distributions only. The remaining calculations are deferred to the appendix. For Poisson random variables, if $G_t(x) = (1+it)^x$ then since $|1+it|^2 = 1+t^2$ the second claim follows. For the first:

$$\mathbb{E}[G_t(X)] = \exp(\lambda((1+it)-1)) = z^\lambda.$$

■

2.2. Variational Distance Between Mixtures

We crucially use the following lemma.

Lemma 5 (Borwein and Erdélyi (1997)) *Let $a_0, a_1, a_2, \cdots \in \{0, 1, -1\}$ be such that not all of them are zero. For any complex number z, let $A(z) \equiv \sum_k a_k z^k$. Then, for some absolute constant c,*

$$\max_{-\pi/L \le t \le \pi/L} |A(e^{it})| \ge e^{-cL} .$$

We will also need the following 'tail bound' lemma.

Lemma 6 *Suppose $a > 1$ is any real number and $r \in \mathbb{R}_+$. For any discrete random variable X with support $\mathbb{Z}$ and pmf f,*

$$\sum_{x \ge r} a^x f(x) \le \frac{\mathbb{E}[a^{2X}]}{a^{r-1}}.$$

Proof Note that, $\Pr(X \ge x) = \Pr(a^{2X-2x} \ge 1) \le \mathbb{E}[a^{2X-2x}]$. We have,

$$\sum_{x \ge r} a^x \Pr(X = x) \le \sum_{x \ge r} a^x \Pr(X \ge x) \le \sum_{x \ge r} a^x \mathbb{E}[a^{2X-2x}] = \mathbb{E}[a^{2X}] \sum_{x \ge r} a^{-x} \le \frac{\mathbb{E}[a^{2X}]}{a^{r-1}}.$$

■

Theorem 7 (TV Lower Bounds) *The following bounds hold on distance between two different mixtures assuming all k parameters are distinct for each mixture.*

- Gaussian: $\mathcal{M} = \frac{1}{k}\sum_{i=1}^k \mathcal{N}(\mu_i, \sigma)$ *and* $\mathcal{M}' = \frac{1}{k}\sum_{i=1}^k \mathcal{N}(\mu_i', \sigma)$ *where* $\mu_i, \mu_i' \in \epsilon\mathbb{Z}$. *Then*
$$\left\|\mathcal{M}' - \mathcal{M}\right\|_{TV} \ge k^{-1} \exp(-\Omega((\sigma/\epsilon)^{2/3})) .$$
- Poisson: $\mathcal{M} = \frac{1}{k}\sum_{i=1}^k Poi(\lambda_i)$ *and* $\mathcal{M}' = \frac{1}{k}\sum_{i=1}^k Poi(\lambda_i')$ *where* $\lambda_i, \lambda_i' \in \{0, 1, \ldots, N\}$. *Then*
$$\left\|\mathcal{M}' - \mathcal{M}\right\|_{TV} \ge k^{-1} \exp(-\Omega(N^{1/3})) .$$
- Chi-Squared: $\mathcal{M} = \frac{1}{k}\sum_{i=1}^k \chi^2(\ell_i)$ *and* $\mathcal{M}' = \frac{1}{k}\sum_{i=1}^k \chi^2(\ell_i')$ *where* $\ell_i, \ell_i' \in \{1, 2, \ldots, N\}$. *Then*
$$\left\|\mathcal{M}' - \mathcal{M}\right\|_{TV} \ge k^{-1} \exp(-\Omega(N^{1/3})) .$$
- Negative Binomial: $\mathcal{M} = \frac{1}{k}\sum_{i=1}^k NB(r_i, p)$ *and* $\mathcal{M}' = \frac{1}{k}\sum_{i=1}^k NB(r_i', p)$ *where* $r_i, r_i' \in \{1, 2, \ldots, N\}$. *Then*
$$\left\|\mathcal{M}' - \mathcal{M}\right\|_{TV} \ge k^{-1} \exp(-\Omega((N/p)^{1/3})) .$$

Proof As above we give the argument for Poisson random variables, deferring the others to the appendix. Let $X \sim \mathcal{M}$ and $X' \sim \mathcal{M}'$. Then, for $w = 1 + it$, from Lemma 4,

$$\mathbb{E}(w^X) - \mathbb{E}(w^{X'}) = \frac{1}{k}\sum_{j=1}^k (e^{it\lambda_j} - e^{it\lambda_j'}).$$

Now we use Lemma 3 with $G(x) = w^x$, $\Omega' = \{0, 1, \ldots, 2N\}$ and $t \leq 1$, to have,

$$\begin{aligned}|\mathbb{E}(w^X) - \mathbb{E}(w^{X'})| &= \Big|\sum_x (w^x \mathcal{M}(x) - w^x \mathcal{M}'(x))\Big| \leq \sum_x |w|^x |\mathcal{M}(x) - \mathcal{M}'(x)| \\ &\leq (1+t^2)^{2N} \sum_x |\mathcal{M}(x) - \mathcal{M}'(x)| + \sum_{x>4N} (1+t^2)^{x/2} e^{-N} \frac{N^x}{x!} \\ &\leq (1+t^2)^{2N} \sum_x |\mathcal{M}(x) - \mathcal{M}'(x)| + \sum_{x>4N} 2^{x/2} e^{-N} \frac{N^x}{x!}.\end{aligned}$$

Now using Lemma 6,

$$\begin{aligned}|\mathbb{E}(w^X) - \mathbb{E}(w^{X'})| &\leq 2(1+t^2)^{2N} \left\|\mathcal{M} - \mathcal{M}'\right\|_{\text{TV}} + \frac{\mathbb{E}[2^X]}{2^{2N-1/2}} \leq 2e^{2t^2 N} \left\|\mathcal{M} - \mathcal{M}'\right\|_{\text{TV}} + \frac{e^N}{2^{2N-1/2}} \\ &= 2e^{2\pi^2 N/L^2} \left\|\mathcal{M} - \mathcal{M}'\right\|_{\text{TV}} + \exp(-\Omega(N)),\end{aligned}$$

by taking $|t| \leq \frac{\pi}{L}$. Now using Lemma 5, there exist an absolute constant c such that,

$$\max_{-\frac{\pi}{L} \leq t \leq \frac{\pi}{L}} \Big|\sum_{j=1}^{k} (e^{it\lambda_j} - e^{it\lambda'_j})\Big| \geq e^{-cL}.$$

Therefore by setting $L = N^{1/3}$,

$$\left\|\mathcal{M} - \mathcal{M}'\right\|_{\text{TV}} \geq (2k)^{-1} e^{-cL - 2\pi^2 N/L^2} - \exp(-\Omega(N)) \geq k^{-1} \exp(-\Omega(N^{1/3})).$$

■

2.3. Parameter Learning

Union Bound Approach for Discrete Distributions We begin with the following proposition which follows from Theorem 7.1 of Devroye and Lugosi (2012).

Lemma 8 *Suppose $F = \{f_\nu\}_{\nu \in \Theta}$ is a class of distribution such that for any $\nu, \nu' \in \Theta$, $\|f_\nu - f_{\nu'}\|_{TV} \geq \delta$. Then $O(\log |\Theta|/\delta^2)$ samples from a distribution f in F suffice to distinguish it from all other distributions in F with high probability.*

For the mixture of Poissons, $\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k} \text{Poi}(\lambda_i)$ where $\lambda_i \in \{0, 1, \ldots, N\}$, the number of choices for parameters in the mixture is $(N+1)^k$. Now using Lemmas 7 and 8, $\exp(O(N^{1/3}))$ samples are sufficient to learn the parameters of the mixture.

Exactly the same argument applies to mixtures of Chi-Squared and Negative-Binomial distributions, yielding $\exp(O(N^{1/3}))$ and $\exp(O((N/p)^{1/3}))$ samples suffice, respectively. However, for Gaussians we need a more intricate approach.

VC Approach for Gaussians To learn the parameters of a Gaussian mixture

$$\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k}\mathcal{N}(\mu_i, \sigma) \quad \text{where} \quad \mu_i \in \{\ldots, -2\epsilon, -\epsilon, 0, \epsilon, 2\epsilon \ldots\}$$

we use the minimum distance estimator precisely defined in (Devroye and Lugosi, 2012, Section 6.8). Let $\mathcal{A} \equiv \{\{x : \mathcal{M}(x) \geq \mathcal{M}'(x)\} : \text{ for any two mixtures } \mathcal{M} \neq \mathcal{M}'\}$ be a collection of subsets. Let P_m denote the empirical probability measure induced by the m samples. Then, choose a mixture $\hat{\mathcal{M}}$ for which the quantity $\sup_{A\in\mathcal{A}} |\Pr_{\sim\hat{\mathcal{M}}}(A) - P_m(A)|$ is minimum (or within $1/m$ of the infimum). This is the minimum distance estimator, whose performance is guaranteed by the following proposition (Devroye and Lugosi, 2012, Thm. 6.4).

Proposition 9 *Given m samples from $\mathcal{M}$ and with $\Delta = \sup_{A\in\mathcal{A}} |\Pr_{\sim\mathcal{M}}(A) - P_m(A)|$, we have*

$$\left\|\hat{\mathcal{M}} - \mathcal{M}\right\|_{TV} \leq 4\Delta + \frac{3}{m}.$$

We now upper bound the right-hand side of the above inequality. Via McDiarmid's inequality and a standard symmetrization argument, Δ is concentrated around its mean which is a function of $VC(\mathcal{A})$, the VC dimension of the class $\mathcal{A}$, see (Devroye and Lugosi, 2012, Section 4.3):

$$\left\|\hat{\mathcal{M}} - \mathcal{M}\right\|_{\text{TV}} \leq 4\Delta + O(1/m) \leq 4\mathbb{E}_{\sim\mathcal{M}}\Delta + O(1/\sqrt{m}) \leq c\sqrt{\frac{VC(\mathcal{A})}{m}},$$

with high probability, for an absolute constant c. This latter term is bounded by the following.

Lemma 10 *For the class $\mathcal{A}$ defined above, the VC dimension is given by $VC(\mathcal{A}) = O(k)$.*

Proof First of all we show that any element of the set $\mathcal{A}$ can be written as union of at most $4k-1$ intervals in $\mathbb{R}$. For this we use the fact that a linear combination of k Gaussian pdfs $f(x) = \sum_{i=1}^{k}\alpha_i f_i(x)$ where f_is normal pdf $\mathcal{N}(\mu_i, \sigma_i^2)$ and $\alpha_i \in \mathbb{R}, 1 \leq i \leq k$ has at most $2k-2$ zero-crossings (Kalai et al., 2012). Therefore, for any two mixtures of interest $\mathcal{M}(x) - \mathcal{M}'(x)$ has at most $4k-2$ zero-crossings. Therefore any $A \in \mathcal{A}$ must be a union of at most $4k-1$ contiguous regions in $\mathbb{R}$. It is now an easy exercise to see that the VC dimension of such a class is $\Theta(k)$. ■

As a result the error of the minimum distance estimator is $O(\sqrt{k/m})$ with high probability. But from Theorem 7, notice that for any other mixture $\mathcal{M}'$ we must have,

$$\left\|\mathcal{M} - \mathcal{M}'\right\|_{\text{TV}} \geq k^{-1}\exp(-\Omega((\sigma/\epsilon)^{2/3})).$$

As long as $\left\|\hat{\mathcal{M}} - \mathcal{M}\right\|_{\text{TV}} \leq \frac{1}{2}\left\|\mathcal{M} - \mathcal{M}'\right\|_{\text{TV}}$ we will exactly identify the parameters. Therefore $m = k^3\exp(O((\sigma/\epsilon)^{2/3}))$ samples suffice to exactly learn the parameters with high probability.

2.4. Extension to Non-Uniform Mixtures

The above results extend to non-uniform mixtures, where the main change is that we require a generalization of Lemma 5. The result, also proved by Borwein and Erdélyi (1997), states that if $a_0, a_1, a_2, \ldots \in [-1, 1]$ with poly(n) precision then $\max_{-\pi/L\leq\theta\leq\pi/L} |A(e^{i\theta})| \geq e^{-cL\log n}$, for an absolute constant c. This weaker bound yields an extra poly(n) factor in the sample complexity.

3. Learning Mixtures via Moments

There are some mixtures where the problem of learning parameters is not amenable to the approach in the previous section. A simple motivating example is learning the parameters $p_i \in \{0, \epsilon, 2\epsilon, 3\epsilon, \ldots, 1\}$ values[3] in the mixture $\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k} \text{Bin}(n, p_i)$. In this section, we present an alternative procedure for learning such mixtures. The basic idea is as follows:

- We compute moments $\mathbb{E}X^\ell$ exactly for $\ell = 0, 1, \ldots, T$ by taking sufficiently many samples. The number of samples will depend on T and the precision of the parameters of the mixture.
- We argue that if T is sufficiently large, then these moments uniquely define the parameters of the mixture. To do this we use a combinatorial result due to Krasikov and Roditty (1997).

In this section, it will be convenient to define a function m_ℓ on multi-sets where

$$m_\ell(A) := \sum_{a \in A} a^\ell \,.$$

Our main result is as follows:

Theorem 11 (Learning Binomial mixtures) *Let $\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k} Bin(n, p_i)$ be a uniform mixture of k binomials, with known shared number of trials n and unknown probabilities $p_1, \ldots, p_k \in \{0, \epsilon, 2\epsilon, \ldots, 1\}$. Then, provided $n \geq 4/\sqrt{\epsilon}$, the first $4/\sqrt{\epsilon}$ moments suffice to learn the parameters p_i and there exists an algorithm that, when given $O(k^2(n/\epsilon)^{8/\sqrt{\epsilon}})$ samples from $\mathcal{M}$, exactly identifies the parameters $\{p_i\}_{i=1}^k$ with high probability.*

Computing the Moments We compute the ℓth moment as $S_{\ell,t} = \sum Y_i^\ell / t$ where $Y_1, \ldots, Y_t \sim X$.

Lemma 12 $\Pr[|S_{\ell,t} - \mathbb{E}X^\ell| \geq \gamma] \leq \frac{\mathbb{E}X^{2\ell}}{t\gamma^2} \leq \frac{(2\ell)!}{\gamma^2 t} \inf_\alpha \left(\frac{\mathbb{E}e^{\alpha X}}{\alpha^{2\ell}}\right)$ *where the last inequality assumes the all the moments of X are non-negative.*

Proof By the Chebyshev bound,

$$\Pr[|S_{\ell,t} - \mathbb{E}X^\ell| \geq \gamma] \leq \frac{Var(S_{\ell,t})}{\gamma^2} = \frac{Var(X^\ell)}{t\gamma^2} \leq \frac{\mathbb{E}X^{2\ell}}{t\gamma^2}.$$

We then use the moment generating function: for all $\alpha > 0$, $\mathbb{E}X^{2\ell} \leq (2\ell)!\mathbb{E}e^{\alpha X}/\alpha^{2\ell}$. ■

The following corollary, tailors the above lemma for a mixture of binomial distributions.

Corollary 13 *If $X \sim \sum_{i=1}^{k} \text{Bin}(n, p_i)/k$ then $\Pr[|S_{\ell,t} - \mathbb{E}X^\ell| \geq \gamma] = \gamma^{-2}n^{2\ell}/t$.*

Fixing n, the ℓ^{th} moment of a mixture of binomial distributions $X \sim \sum_{i=1}^{k} \text{Bin}(n, p_i)/k$ is

$$\mathbb{E}X^\ell = \sum_{i=1}^{k} f(p_i)/k$$

where f is a polynomial of degree at most ℓ with integer coefficients (Belkin and Sinha, 2010). If p_i is an integer multiple of ϵ then this implies $k(\mathbb{E}X^\ell)/\epsilon^\ell$ is integral and therefore any mixture with a different ℓth moment differs by at least ϵ^ℓ/k. Hence, learning the ℓth moment up to $\gamma_\ell < \epsilon^\ell/(2k)$ implies learning the moment exactly.

3. Note that we are implicitly assuming $1/\epsilon$ is integral here and henceforth.

Lemma 14 *For $X \sim \text{Bin}(n, p)$, $\mathbb{E}X^\ell$ is a polynomial in p of degree exactly ℓ if $n \geq \ell$.*

The proof of the lemma is relegated to the appendix.

Theorem 15 *$O(k^2(n/\epsilon)^{8/\sqrt{\epsilon}})$ samples are sufficient to exactly learn the first $4/\sqrt{\epsilon}$ moments of a uniform mixture of k binomial distributions $\sum_{i=1}^k \text{Bin}(n, p_i)/k$ with probability at least $7/8$ where each $p_i \in \{0, \epsilon, 2\epsilon, \ldots, 1\}$.*

Proof Let $T = 4/\sqrt{\epsilon}$. From Corollary 20 and the preceding discussion, learning the ℓth moment exactly with failure probability $1/9^{1+T-\ell}$ requires

$$t = \gamma_\ell^{-2} n^{2\ell} 9^{1+T-\ell} = O(k^2 9^{1+T-\ell} n^{2\ell}/\epsilon^{2\ell}) = O(k^2 9^T (n/3\epsilon)^{2\ell})$$

samples. And hence, we can compute all ℓth moments exactly for $1 \leq \ell \leq 4/\sqrt{\epsilon}$ using

$$\sum_{\ell=1}^{T} O(k^2 9^T (n/3\epsilon)^{2\ell}) = O(k^2 (n/\epsilon)^{2T})$$

samples with failure probability $\sum_{\ell=1}^T 1/9^{1+T-\ell} < \sum_{i=1}^\infty 1/9^i = 1/8$. ■

How many moments determine the parameters It remains to show the first $4/\sqrt{\epsilon}$ moments suffice to determine the p_i values in the mixture $X \sim \sum_{i=1}^k \text{Bin}(n, p_i)/k$ provided $n \geq \frac{4}{\epsilon}$. To do this suppose there exists another mixture $Y \sim \sum_{i=1}^k \text{Bin}(n, q_i)/k$ and we will argue that

$$\mathbb{E}X^\ell = \mathbb{E}Y^\ell \text{ for } \ell = 0, 1, \ldots, 4\sqrt{1/\epsilon}$$

implies $\{p_i\}_{i\in[k]} = \{q_i\}_{i\in[k]}$. To argue this, define integers $\alpha_i, \beta_i \in \{0, 1, \ldots, 1/\epsilon\}$ such at that $p_i = \alpha_i \epsilon$ and $q_i = \beta_i \epsilon$. Let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_k\}$ and $\mathcal{B} = \{\beta_1, \ldots, \beta_k\}$. Then,

$$\mathbb{E}X = \mathbb{E}Y \implies \sum_i \alpha_i = \sum_i \beta_i \implies m_1(\mathcal{A}) = m_1(\mathcal{B})$$

and, after some algebraic manipulation, it can be shown that for all $\ell \in \{2, 3, \ldots\}$,

$$\left(\forall \ell' \in \{0, 1, \ldots, \ell - 1\}\ ,\ \sum_i \alpha_i^{\ell'} = \sum_i \beta_i^{\ell'}\right) \text{ and } \mathbb{E}X^\ell = \mathbb{E}Y^\ell$$
$$\implies \left(\sum_i \alpha_i^\ell = \sum_i \beta_i^\ell\right) \implies m_\ell(\mathcal{A}) = m_\ell(\mathcal{B})\ .$$

Hence, if the first T moments match $m_\ell(\mathcal{A}) = m_\ell(\mathcal{B})$ for all $\ell = 0, 1, \ldots, T$. But the following theorem establishes that if $T = 4\sqrt{1/\epsilon}$ then this implies $\mathcal{A} = \mathcal{B}$.

Theorem 16 (Krasikov and Roditty (1997)) *For any two subsets S, T of $\{0, 1, \ldots, n-1\}$, then*

$$S = T \text{ iff } \left(m_k(S) = m_k(T) \text{ for all } k = 0, 1, \ldots, 4\sqrt{n}\right)\ .$$

We note that the above theorem is essentially tight. Specifically, there exists $S \neq T$ with $m_k(S) = m_k(T)$ for $k = 0, 1, \ldots, cn/\log n$ for some c. As a consequence of this, we note that even the exact values of the $c\sqrt{n}/\log n$ moments are insufficient to learn the parameters of the distribution. For an example in terms of Gaussian mixtures, even given the promise $\mu_i \in \{0, 1, \ldots, n-1\}$ are distinct, then the first $c\sqrt{n}/\log n$ moments of $\sum_i \mathcal{N}(\mu_i, 1)$ are insufficient to uniquely determine μ_i whereas the first $4\sqrt{n}$ moments are sufficient.

3.1. Extension to Non-Uniform Distributions

We now consider extending the framework to non-uniform distributions. In this case, the method of computing the moments is identical to the uniform case. However, when arguing that a small number of moments suffices we can no longer appeal to the Theorem 16.

To handle non-uniform distribution we introduce a precision variable q and assume that the weights of the component distributions $\omega_1, \omega_2, \ldots, \omega_k$ are of the form:

$$\omega_i = \frac{w_i}{\sum_{i=1}^{k} w_i}$$

where $w_i \in \{0, 1, \ldots, q-1\}$. Then, in the above framework if we are trying to learn parameters $\alpha_1, \ldots, \alpha_k$ then the moments are going to define a multi-set consisting of w_i copies of α_i for each $i \in [k]$. To quantify how many moments suffice in this case, we need to prove a variant of Theorem 16. The proof is a relatively straight-forward generalization of proof by Scott (1997) and can be found in the appendix.

Theorem 17 *For any two multi-sets S, T where each element is in $\{0, 1, \ldots, n-1\}$ and the multiplicity of each element is at most $q-1$, then $S = T$ if and only if $m_k(S) = m_k(T)$ for all $k = 0, 1, \ldots, 2\sqrt{qn \log qn}$.*

Acknowledgements The work was partially supported by NSF grants CCF-1909046, CCF-1934846, CCF-1908849, and CCF-1637536.

References

Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Conference on Learning Theory*, 2005.

Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Symposium on Theory of Computing*, 2001.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science*, 2010.

P. Borwein and T. Erdélyi. Littlewood-type problems on subarcs of the unit circle. *Indiana University Mathematics Journal*, 1997.

Peter Borwein. *The Prouhet—Tarry—Escott Problem*, pages 85–95. Springer New York, New York, NY, 2002. ISBN 978-0-387-21652-2. doi: 10.1007/978-0-387-21652-2_11. URL https://doi.org/10.1007/978-0-387-21652-2_11.

Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Symposium on Discrete Algorithms*, 2013.

Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 2014.

Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science*, pages 634–644, 1999.

Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Conference on Learning Theory*, 2014.

Anindya De, Ryan O'Donnell, and Rocco A. Servedio. Optimal mean-based algorithms for trace reconstruction. In *Symposium on Theory of Computing*, 2017a.

Anindya De, Ryan O'Donnell, and Rocco A. Servedio. Sharp bounds for population recovery. *CoRR*, abs/1703.01474, 2017b. URL http://arxiv.org/abs/1703.01474.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Ilias Diakonikolas. Learning structured distributions. *Handbook of Big Data*, 267, 2016.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional gaussians and gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 73–84. IEEE, 2017.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060. ACM, 2018.

Jon Feldman, Ryan O'Donnell, and Rocco A Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 2008.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Symposium on Theory of Computing*, 2015.

Godfrey Harold Hardy, Edward Maitland Wright, et al. *An introduction to the theory of numbers*. Oxford university press, 1979.

Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Symposium on Theory of Computing*, 2018.

Adam Kalai, Ankur Moitra, and Gregory Valiant. Disentangling Gaussians. *Communications of the ACM*, 55(2):113–120, 2012.

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046. ACM, 2018.

I. Krasikov and Y. Roditty. On a reconstruction problem for sequences. *Journal of Combinatorial Theory, Series A*, 1997.

Akshay Krishnamurthy, Arya Mazumdar, Andrew McGregor, and Soumyabrata Pal. Trace reconstruction: Generalized and parameterized. In *27th Annual European Symposium on Algorithms, ESA 2019, September 9-11, 2019, Munich/Garching, Germany.*, pages 68:1–68:25, 2019.

Ankur Moitra. *Algorithmic aspects of machine learning*. Cambridge University Press, 2018.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science*, 2010.

Fedor Nazarov and Yuval Peres. Trace reconstruction with $\exp(O(n^{1/3})$ samples. In *Symposium on Theory of Computing*, 2017.

Alex D. Scott. Reconstructing sequences. *Discrete Mathematics*, 1997.

Ananda Theertha Suresh, Alon Orlitsky, Jayadev Acharya, and Ashkan Jafarpour. Near-optimal-sample estimators for spherical gaussian mixtures. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2014.

D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley, 1985.

Eric W. Weisstein. Geometric distribution. *From MathWorld–A Wolfram Web Resource*, 2019. URL http://mathworld.wolfram.com/GeometricDistribution.html.

4. Omitted Proofs

Additional calculations for Lemma 4. We consider each distribution in turn:

- *Gaussian:* Observe that $\mathbb{E}[G_t(X)]$ is precisely the characteristic function. Clearly we have $\|G_t\|_\infty = 1$ and further
$$\mathbb{E}[G_t(X)] = \exp(it\mu - \sigma^2 t^2/2) = \exp(-\sigma^2 t^2/2) z^\mu.$$
- *Poisson:* If $G_t(x) = (1+it)^x$ then since $|1+it|^2 = 1+t^2$ the second claim follows. For the first:
$$\mathbb{E}[G_t(X)] = \exp(\lambda((1+it)-1)) = z^\lambda.$$
- *Chi-Squared:* Let $w_t = \exp(1/2 - e^{-2it}/2)$ then $|w_t|^2 = |e^{1-e^{-2it}}| = |e^{1-\cos 2t} e^{i \sin 2t}| \leq e^{ct^2 + O(t^4)}$ and
$$\mathbb{E}[G_t(X)] = (1 - 2\ln w_t)^{-\frac{\ell}{2}} = z^\ell.$$
- *Negative Binomial:* Let $w_t = 1/p - (1/p - 1)e^{-it}$ then $|w_t|^2 = \frac{1+(1-p)^2 - 2(1-p)\cos t}{p^2} = \frac{p^2 + 4(1-p)\sin^2(t/2)}{p^2} \leq e^{\frac{(1-p)t^2}{p^2}}$ and
$$\mathbb{E}[G_t(X)] = \left(\frac{1-p}{1-pw_t}\right)^r = z^r.$$

4.1. Additional calculations for Theorem 7.

- *Gaussian:* The characteristic function of a Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$ is

$$C_{\mathcal{N}}(t) = \mathbb{E}e^{itX} = e^{it\mu - \frac{t^2\sigma^2}{2}}.$$

Therefore we have that

$$C_{\mathcal{M}}(t) - C_{\mathcal{M}'}(t) \geq \frac{e^{-\frac{t^2\sigma^2}{2}}}{k} \sum_{j=1}^{k} (e^{it\mu_j} - e^{it\mu'_j}).$$

Now, using Lemma 5, there exist an absolute constant c such that,

$$\max_{-\frac{\pi}{\epsilon L} \leq t \leq \frac{\pi}{\epsilon L}} \Big| \sum_{j=1}^{k} (e^{it\mu_j} - e^{it\mu'_j}) \Big| \geq e^{-cL}.$$

Also, for $t \in (-\frac{\pi}{\epsilon L}, \frac{\pi}{\epsilon L})$, $e^{-\frac{t^2\sigma^2}{2}} \geq e^{-\frac{\sigma^2\pi^2}{2\epsilon^2 L^2}}$. And therefore,

$$\Big| C_{\mathcal{M}}(t) - C_{\mathcal{M}'}(t) \Big| \geq \frac{1}{k} e^{-\frac{\sigma^2\pi^2}{2\epsilon^2 L^2} - cL}.$$

By substituting $L = \frac{(\pi\sigma)^{2/3}}{(\epsilon^2 c)^{1/3}}$ above we conclude that there exists t such that

$$\Big| C_{\mathcal{M}}(t) - C_{\mathcal{M}'}(t) \Big| \geq \frac{1}{k} e^{-\frac{3}{2}(c\pi\sigma/\epsilon)^{2/3}}.$$

Now using Lemma 3, we have $\|\mathcal{M}' - \mathcal{M}\|_{\text{TV}} \geq k^{-1} \exp(-\Omega((\sigma/\epsilon)^{2/3}))$.

- *Chi-Squared:* Let $X \sim \mathcal{M}$ and $X' \sim \mathcal{M}'$. Then, for $w = \exp(1/2 - e^{-2it}/2)$, from Lemma 4,

$$\mathbb{E}(w^X) - \mathbb{E}(w^{X'}) = \frac{1}{k} \sum_{j=1}^{k} (e^{it\ell_j} - e^{it\ell'_j}).$$

Now we use Lemma 3, with $\Omega' = [0, 2N]$ we have,

$$\|\mathcal{M} - \mathcal{M}'\|_{TV} \geq e^{-2ct^2N} \Big(\Big| \mathbb{E}(w^X) - \mathbb{E}(w^{X'}) \Big| - \int_{x>2N} \exp(ct^2x) f(x) dx \Big),$$

where $f \sim \chi^2(N)$. We have,

$$\begin{aligned} \int_{x>2N} \exp(ct^2x) f(x) dx &= \frac{1}{(1-2ct^2)^{N/2-1}} \int_{y>2N(1-2ct^2)} f(y) dy \leq \frac{e^{-N(1-4ct^2)^2/8}}{(1-2ct^2)^{N/2-1}} \\ &\leq \exp(-\Omega(N)), \end{aligned}$$

where we have used the pdf of chi-squared distribution and the tail bounds for chi-squared. Now using Lemma 5, and taking $|t| \leq \frac{\pi}{L}$,

$$\|\mathcal{M} - \mathcal{M}'\|_{TV} \geq k^{-1} e^{-c'L - 2ct^2N} - \exp(-\Omega(n)) \geq k^{-1} \exp(-c'L - 2\pi^2 N/L^2) - \exp(-\Omega(N)).$$

Again setting, $L = N^{1/3}$,

$$\|\mathcal{M} - \mathcal{M}'\|_{TV} \geq k^{-1} \exp(-\Omega(N^{1/3})).$$

- *Negative-Binomial:* Let $X \sim \mathcal{M}$ and $X' \sim \mathcal{M}'$. Then, for $w = 1/p - (1/p - 1)e^{-it}$, from Lemma 4, taking $G(x) = w^x$,

$$\mathbb{E}(w^X) - \mathbb{E}(w^{X'}) = \frac{1}{k}\sum_{j=1}^{k}(e^{itr_j} - e^{itr'_j}).$$

Now we use Lemma 3, with $\Omega' = [0, 6pN/(1-p)]$ we have,

$$\|\mathcal{M} - \mathcal{M}'\|_{TV} \geq e^{-12ct^2N/p}\Big(\Big|\mathbb{E}(w^X) - \mathbb{E}(w^{X'})\Big| - \sum_{x > \frac{6Np}{1-p}} |w|^x u(x)\Big),$$

where $u(x) = \binom{x+N-1}{x}(1-p)^N p^x$. We have $|w| \leq e^{c(1-p)t^2/p^2} \leq e^{c(1-p)/p^2}$ for $t < 1$. Using Lemma 6, with $X \sim NB(N, p)$, we have,

$$\sum_{x > \frac{6Np}{1-p}} \exp(cx(1-p)/p^2)u(x) \leq a^{1-\frac{6Np}{1-p}}\mathbb{E}[a^{2X}] = a^{1-\frac{6Np}{1-p}}\Big(\frac{1-p}{1-pa^2}\Big)^N = \exp(-\Omega(N)),$$

where, $a = \exp(c(1-p)/p^2) > 1$. Now using Lemma 5, and taking $|t| \leq \frac{\pi}{L}$,

$$\begin{aligned}\|\mathcal{M} - \mathcal{M}'\|_{TV} &\geq k^{-1}e^{-c'L - 12ct^2N/p} - \exp(-\Omega(n)) \\ &\geq k^{-1}\exp(-c'L - 12\pi^2 N/(pL^2)) - \exp(-\Omega(N)).\end{aligned}$$

Setting $L = (N/p)^{1/3}$,

$$\|\mathcal{M} - \mathcal{M}'\|_{TV} \geq k^{-1}\exp(-\Omega((N/p)^{1/3})).$$

4.2. Proof of Theorem 17

Let $\mathbf{a}$ be the characteristic vector of a subset $S \subset \mathcal{U}$. Let $s_\ell = m_\ell(S)$ on this set and let $\mathbf{s} = (s_0, s_1, \ldots, s_{k-1})$. We need to prove $\mathbf{a}$ is uniquely determined by $\mathbf{s}$.

Let us define

$$n_{i,p}(\mathbf{a}) := \sum_{r \equiv_p i} a_r \pmod{p}.$$

Claim 1 *For a prime number p and $i \not\equiv_p 0$, we have $n_{i,p}(\mathbf{a}) \equiv_p s_0 - \sum_j \binom{p-1}{j} s_j (-i)^{p-1-j}$*

Proof

$$n_{i,p}(\mathbf{a}) = \sum_{r \equiv_p i} a_r \pmod{p}$$

Recall that Fermat's theorem (Hardy et al. (1979)) says that for any prime p and any number $\alpha \not\equiv_p 0$, we must have that $\alpha^{p-1} \equiv_p 1$. Hence, for a prime number p and some number $i \not\equiv_p 0$, we have

$$\begin{aligned} s_0(\mathbf{a}) - \sum_j \binom{p-1}{j} s_j (-i)^{p-1-j} &\equiv_p \sum_r a_r - \sum_j \binom{p-1}{j} \sum_r a_r r^j (-i)^{p-1-j} \\ &\equiv_p \sum_r a_r - \sum_r a_r \sum_j \binom{p-1}{j} r^j (-i)^{p-1-j} \\ &\equiv_p \sum_r a_r - \sum_r a_r (r-i)^{p-1} \\ &\equiv_p \sum_{r \equiv_p i} a_r \equiv_p n_{i,p}(\mathbf{a}) \ . \end{aligned}$$

■

Since the value of $n_{i,p}$ is at most $\lceil qn/p \rceil$, we can obtain the value of $n_{i,p}$ exactly if p is chosen to be greater than $\sqrt{qn}$. Now, let us denote the vector $\mathbf{v}_{i,p} \in \mathbb{F}_q^n$ where the ℓth entry is

$$v_{i,p}[\ell] = \begin{cases} 1 & \text{if } \ell \equiv_p i \\ 0 & \text{otherwise} \end{cases} \ .$$

Therefore, consider two different subsets $S, S' \subset \mathcal{U}$ and assume that their characteristic vectors are $\mathbf{a}$ and $\mathbf{b}$ respectively. Therefore, if $\mathbf{a}$ and $\mathbf{b}$ both give rise to the same value of $\mathbf{s}$, then $\mathbf{a}.\mathbf{v}_{i,p} = \mathbf{b}.\mathbf{v}_{i,p}$. Hence, if the set of vectors

$$\mathcal{S} = \{v_{i,p} \mid \sqrt{qn} \leq p \leq k, 0 \leq i \leq p-1, p \text{ prime}\}$$

spans $\mathbb{F}_q^n$, then it must imply that $\mathbf{a} = \mathbf{b}$ and our proof will be complete. Consider a subset $\mathcal{T} \subset \mathcal{S}$ defined by

$$\mathcal{T} = \{v_{i,p} \mid \sqrt{qn} \leq p \leq k, 1 \leq i \leq p-1, p \text{ prime}\}$$

Now, there are two possible cases. First, let us assume that the vectors in $\mathcal{T}$ are not all linearly independent in $\mathbb{F}_q$. In that case, we must have a set of tuples $(i_1, p_1), (i_2, p_2), \ldots, (i_m, p_m)$ such that

$$\sum_{j=1}^{m} \alpha_j \mathbf{v}_{(i_j, p_j)} \equiv_q 0 \tag{1}$$

where $0 \neq \alpha_j \in \mathbb{F}_q$ for all j. Now, by the Chinese Remainder Theorem, we can find an integer r such that $r \equiv_{p_1} i_1$ and $r \equiv_{p_j} 0$ for all $p_j \neq p_1$. Define an infinite dimensional vector $\tilde{\mathbf{v}}$ where the ℓth entry is

$$\tilde{\mathbf{v}}[\ell] = \sum_{j=1}^{m} \alpha_j \mathbb{1}\left[\ell \equiv_{p_j} i_j\right]$$

Since, $i_j \not\equiv_{p_j} 0$, it is evident that $\tilde{\mathbf{v}}[r] \not\equiv_q 0$ Now, let s be the smallest number such that $\tilde{\mathbf{v}}[s] \neq 0$ and $s > n$ because of our assumption in Eq. 1. Now consider the vector $\mathbf{v}_t$ where

$$\mathbf{v}_t = \sum_{j=1}^{m} \alpha_j \mathbf{v}_{i_j - s + t, p_j}$$

Now, $\mathbf{v}_t^i = 0$ for all $i < t$ and $\mathbf{v}_t^t \neq 0$. Hence, the set $\{\mathbf{v}_t\}_{t=1}^n$ are in the span of $\mathcal{S}$ and also span $\mathbb{F}_q^n$.

For the second case, let us assume that the vectors in $\mathcal{T}$ are linearly independent. We require the size of $\mathcal{T} > n$ so that the vectors in $\mathcal{T}$ span $\mathbb{F}_q^n$. From the prime number theorem we know that

$$\sum_{p \text{ prime}: p < x} p \sim \frac{x^2}{2 \log x}$$

and hence we simply need that

$$\frac{k^2}{2 \log k} - \frac{qn}{\log n} > n \, .$$

Therefore, $k > (1 + o(1))\sqrt{qn \log qn}$ is sufficient.

4.3. Algebraic method for Geometric distribution

We will denote the Geometric distribution with success parameter $0 < p < 1$ as Geo(p) and it has the following form: for a random variable X distributed according to Geo(p), $\Pr(X = x) = (1-p)^x p$ where $x \in \{0, 1, 2, \dots\}$.

Theorem 18 (Learning mixtures of Geometric Distribution) *Let $\mathcal{M} = \frac{1}{k} \sum_{i=1}^k Geo(p_i)$ be a uniform mixture of k Geometric distributions, with unknown probabilities*

$$p_1, \dots, p_k \in \{\frac{1}{1+n\epsilon}, \frac{1}{1+(n-1)\epsilon}, \dots, 1\} \, .$$

Then, the first $4\sqrt{n}$ moments suffice to learn the parameters p_i and there exists an algorithm that, when given $O\left(k^2 \left(\frac{\sqrt{n}}{\epsilon}\right)^{8\sqrt{n}}\right)$ samples from $\mathcal{M}$, exactly identifies the parameters $\{p_i\}_{i=1}^k$ with high probability.

Computing the moments. We compute the ℓth moment in the natural way again. Let $Y_1, \dots, Y_t \sim X$ and let

$$S_\ell = \sum Y_i^\ell / t \, .$$

Lemma 19 (Restating Lemma 12) $\Pr[|S_\ell - \mathbb{E}X^\ell| \geq \gamma] \leq \frac{\mathbb{E}X^{2\ell}}{t\gamma^2} \leq \frac{(2\ell)!}{\gamma^2 t} \inf_\alpha \left(\frac{\mathbb{E}e^{\alpha X}}{\alpha^{2\ell}}\right)$ *where the last inequality assumes the all the moments of X are non-negative.*

The following corollary, tailors the above lemma for a mixture of geometric distributions.

Corollary 20 *If $X \sim \sum_{i=1}^k \text{Geo}(p_i)/k$ then $\Pr[|S_\ell - \mathbb{E}X^\ell| \geq \gamma] \leq \frac{2}{t\gamma^2} \left(\frac{4\ell}{\min_i p_i}\right)^{2\ell+1}$.*

Proof Given a random variable $Z \sim \text{Geo}(p)$, we will show that $\mathbb{E}Z^k \leq 2\left(\frac{2k}{p}\right)^{k+1}$ for all integer valued $k \geq 0$. It is known that (Weisstein, 2019)

$$\mathbb{E}Z^k = p\text{Li}_{-k}(1-p)$$

where $\text{Li}_{-k}(z)$ is the polylogarithmic function of order $-k$ and argument z, defined explicitly as

$$\text{Li}_{-k}(1-p) = \frac{1}{p^{k+1}} \sum_{j=0}^{k-1} \left\langle {k \atop j} \right\rangle (1-p)^{k-j}$$

with $\left\langle {k \atop j} \right\rangle$ being the Eulerian numbers (see below). Hence, it can be observed that $\mathbb{E}Z^k$ is a polynomial in $\frac{1}{p}$ of degree k. Denoting $C_k = \max_{0 \leq j \leq k-1} \left\langle {k \atop j} \right\rangle$ and substituting it, we get that

$$\mathbb{E}Z^k \leq \frac{C_k}{p^k} \sum_{j=0}^{k-1} (1-p)^{k-j} = C_k \left(\frac{1}{p} - 1\right)\left(\frac{1}{p^k} - \left(\frac{1}{p} - 1\right)^k\right) < \frac{2C_k}{p^{k+1}}.$$

From the definition of Eulerian numbers, we can also see that

$$\left\langle {k \atop j} \right\rangle = \sum_{t=0}^{j} (-1)^t \binom{k+1}{t} (j+1-t)^k \leq (j+1)^k 2^{k+1} < (2k)^{k+1}.$$

Putting everything together and by appealing to Lemma 12, we get the statement of the corollary. ■

For the geometric distribution,

$$\mathbb{E}X^\ell = \sum_{i=1}^{k} f(1/p_i)/k$$

where f is a degree ℓ polynomial with integer coefficients. If $1/p_i - 1$ is an integer multiple of ϵ then this implies $k(\mathbb{E}X^\ell)/\epsilon^\ell$ is integral and therefore any mixture with a different ℓth moment must differ by at least ϵ^ℓ/k. Hence, learning the ℓth moment up to $\gamma_\ell < \epsilon^\ell/(2k)$ implies learning the moment exactly.

Lemma 21 *$O\left(k^2\left(\frac{\sqrt{n}}{\epsilon}\right)^{8\sqrt{n}}\right)$ samples are sufficient to exactly learn the first $4\sqrt{n}$ moments of a uniform mixture of k Geometric distributions $\sum_{i=1}^{k} \text{Geo}(p_i)/k$ with probability at least $7/8$ where each $\frac{1}{p_i} \in \{1, 1+\epsilon, 1+2\epsilon, \ldots, 1+n\epsilon\}$.*

Proof Let $T = 4\sqrt{n}$. From Corollary 20 and the preceding discussion, learning the ℓth moment exactly with failure probability $1/9^{1+T-\ell}$ requires

$$t = \gamma_\ell^{-2} 2\left(4\ell\right)^{2\ell+1} 9^{1+T-\ell} = O\left(k^2 9^{1+T-\ell} \ell^{2\ell}/\epsilon^{2\ell}\right) = O\left(k^2 9^T \left(\frac{\ell}{3\epsilon}\right)^{2\ell}\right)$$

samples. And hence, we can compute all ℓth moments exactly for $1 \leq \ell \leq 4\sqrt{n}$ using

$$\sum_{\ell=1}^{T} O\left(k^2 9^T \left(\frac{\ell}{3\epsilon}\right)^{2\ell}\right) = O\left(k^2 \left(\frac{T}{\epsilon}\right)^{2T}\right)$$

samples with failure probability $\sum_{\ell=1}^{T} 1/9^{1+T-\ell} < \sum_{i=1}^{\infty} 1/9^i = 1/8$. ■

How many moments needed to determine the parameters? It remains to show the first $4\sqrt{n}$ moments suffice to determine the p_i values in the mixture $X \sim \sum_{i=1}^{k} \text{Geo}(p_i)/k$. To do this suppose there exists another mixture $Y \sim \sum_{i=1}^{k} \text{Geo}(q_i)/k$ and we will argue that

$$\mathbb{E}X^\ell = \mathbb{E}Y^\ell \text{ for } \ell = 0, 1, \ldots, 4\sqrt{n}$$

implies $\{p_i\}_{i\in[k]} = \{q_i\}_{i\in[k]}$. To argue this, define integers $\alpha_i, \beta_i \in \{0, 1, \ldots, n\}$ such that $p_i = \frac{1}{1+\alpha_i\epsilon}$ and $q_i = \frac{1}{1+\beta_i\epsilon}$. Let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_k\}$ and $\mathcal{B} = \{\beta_1, \ldots, \beta_k\}$. Then,

$$\mathbb{E}X = \mathbb{E}Y \implies \sum_i 1/p_i = \sum_i 1/q_i \implies \sum_i \alpha_i = \sum_i \beta_i \implies m_1(\mathcal{A}) = m_1(\mathcal{B})$$

and, after some algebraic manipulation, it can be shown that for all $\ell \in \{2, 3, \ldots\}$,

$$\left(\forall \ell' \in \{0, 1, \ldots, \ell-1\}\ ,\ \sum_i \alpha_i^{\ell'} = \sum_i \beta_i^{\ell'}\right) \text{ and } \mathbb{E}X^\ell = \mathbb{E}Y^\ell \implies \sum_i \alpha_i^\ell = \sum_i \beta_i^\ell$$
$$\implies m_\ell(\mathcal{A}) = m_\ell(\mathcal{B})\ .$$

Hence, if the first T moments match, $m_\ell(\mathcal{A}) = m_\ell(\mathcal{B})$ for all $\ell = 0, 1, \ldots, T$. But, again Theorem 16 establishes that if $T = 4\sqrt{n}$ then this implies $\mathcal{A} = \mathcal{B}$.

Alternative Technique. In the previous analysis the parameters of the geometric distribution (p_i's) had to belong to the set $\{1, \frac{1}{1+\epsilon}, \frac{1}{1+2\epsilon}, \ldots, \frac{1}{1+n\epsilon}\}$. The reason we had to choose this set is because the moments were polynomials in inverse of the parameters ($\frac{1}{p_i}$'s). However it is also possible to obtain a sample complexity bound when the parameters belong to the set $\{0, \epsilon, 2\epsilon, \ldots, 1\}$. This can be done by estimating the probability mass function of the mixture at the discrete points $\{0, 1, 2, \ldots\}$. We have the following theorem in this case.

Theorem 22 (Learning mixtures of geometric distributions (alternative)) *Let $\mathcal{M} = \frac{1}{k}\sum_{i=1}^{k} Geo(p_i)$ be a uniform mixture of k geometric distributions, with unknown probabilities $p_1, \ldots, p_k \in \{0, \epsilon, \ldots, 1\}$. Then, the first $4/\sqrt{\epsilon}$ moments suffice to learn the parameters p_i and there exists an algorithm that, when given $O\left(\frac{k^2}{\epsilon^{8/\sqrt{\epsilon}+2}} \log \frac{1}{\epsilon}\right)$ samples from $\mathcal{M}$, exactly identifies the parameters $\{p_i\}_{i=1}^{k}$ with high probability.*

Recall that for a random variable $X \sim \mathcal{M}$ distributed according to the mixture of geometric distributions, we have

$$\Pr(X = 0) = \frac{1}{k}\sum_i p_i$$
$$\Pr(X = 1) = \frac{1}{k}\sum_i p_i - p_i^2$$
$$\Pr(X = 2) = \frac{1}{k}\sum_i p_i - 2p_i^2 + p_i^3$$

and more generally,

$$\Pr(X = k) = \frac{1}{k}\sum_i (1 - p_i)^k p_i$$

which is a polynomial in degree $k+1$. Now, for the mixture $X \sim 1/k \sum_{i=1}^{k} \text{Geo}(p_i)$, we need to argue that estimating the probabilities $\Pr(X=\ell)$ for $\ell = 0, 1, \ldots, 4\sqrt{1/\epsilon}$ is sufficient to recover the parameters p_i. Again, suppose there exists another mixture $Y \sim 1/k \sum_{i=1}^{k} \text{Geo}(q_i)$ such that

$$\Pr(X=\ell) = \Pr(Y=\ell) \text{ for } \ell = 0, 1, \ldots, 4\sqrt{1/\epsilon}$$

and we will argue that this implies $\{p_i\}_{i\in[k]} = \{q_i\}_{i\in[k]}$. As before, define integers $\alpha_i, \beta_i \in \{0, 1, \ldots, \frac{1}{\epsilon}\}$ such that $p_i = \alpha_i\epsilon$ and $q_i = \beta_i\epsilon$. Let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_k\}$ and $\mathcal{B} = \{\beta_1, \ldots, \beta_k\}$ and it can be shown after some algebraic manipulations that

$$\left(\forall \ell' \in \{0, 1, \ldots, \ell-1\}\,,\ \sum_i \alpha_i^{\ell'+1} = \sum_i \beta_i^{\ell'+1}\right) \text{ and } \Pr(X=\ell) = \Pr(Y=\ell)$$
$$\implies \sum_i \alpha_i^{\ell+1} = \sum_i \beta_i^{\ell+1} \implies m_{\ell+1}(\mathcal{A}) = m_{\ell+1}(\mathcal{B}).$$

Notice that $m_0(\mathcal{A}) = m_0(\mathcal{B})$ trivially because both of them contain k components. Again, Theorem 16 establishes that if $m_\ell(\mathcal{A}) = m_\ell(\mathcal{B})$ for $\ell \in \{0, 1, \ldots 4\sqrt{1/\epsilon}\}$ then this implies $\mathcal{A} = \mathcal{B}$.

Computing the probabilities. Suppose $Y_1, Y_2, \ldots, Y_t$ are i.i.d. with $X \sim 1/k \sum_{i=1}^{k} \text{Geo}(p_i)$. Let us denote S_ℓ as the empirical probability that we calculate as,

$$S_\ell = \frac{\sum_{i=1}^{t} \mathbb{1}[Y_i = \ell]}{t}.$$

It is obvious that $\mathbb{E}S_\ell = \Pr(X=\ell)$. Now, using Chernoff bound, we have

$$\Pr(|S_\ell - \Pr(X=\ell)| \geq \gamma_\ell) \leq 2e^{-t\gamma_\ell^2/3}.$$

Again, recall that

$$\Pr(X=\ell) = \sum_i \frac{f(p_i)}{k}$$

where $f(\cdot)$ is a polynomial of degree $\ell+1$ with integer coefficients. If p_i is an integer multiple of ϵ then this implies $kS_\ell/\epsilon^{\ell+1}$ is integral and therefore any mixture with a different ℓth moment has a ℓ moment that differs by at least $\epsilon^{\ell+1}/k$. Hence, learning the ℓth moment up to $\gamma_\ell < \epsilon^{\ell+1}/(2k)$ implies learning the moment exactly. We will use $t = \frac{12k^2}{\epsilon^{8/\sqrt{\epsilon}+2}} \log \frac{64}{\sqrt{\epsilon}}$ number of samples and we will show it will be sufficient to succeed with a probability of at least $\frac{7}{8}$. We will estimate the probabilities as mentioned above and therefore the failure probability can be calculated by using the Chernoff Bound and a union bound over $\frac{4}{\sqrt{\epsilon}}$ probabilities to be estimated. Therefore the probability of failure is bounded above by,

$$\sum_\ell 2\exp(-t\gamma_\ell^2/3) \leq \frac{8}{\sqrt{\epsilon}} \max_\ell \exp(-t\gamma_\ell^2/3) = \frac{8}{\sqrt{\epsilon}} \exp(-t \min_\ell \gamma_\ell^2/3)$$
$$= \frac{8}{\sqrt{\epsilon}} \exp(-t\epsilon^{8/\sqrt{\epsilon}+2}/(12k^2)) \leq \frac{1}{8}$$

and hence the proof is complete.

4.4. Proof of Lemma 14

We will prove that for $X \sim \text{Bin}(n,p)$, the leading term of $\mathbb{E}X^\ell$ is $\prod_{i=0}^{\ell-1}(n-i)p^\ell$. Since for $n \geq \ell$, $\prod_{i=0}^{\ell-1}(n-i) \neq 0$, this implies that $\mathbb{E}X^\ell$ is a polynomial of degree exactly ℓ. We will prove this by induction. Since $X \sim \text{Bin}(n,p)$, we know that $\mathbb{E}X = np$. This verifies the base case. Now, in the induction step, let us assume that the leading term of $\mathbb{E}X^k$ is $\prod_{i=0}^{k-1}(n-i)p^k$. It is known that (see Belkin and Sinha (2010))

$$\mathbb{E}X^{k+1} = np\mathbb{E}X^k + p(1-p)\frac{d\mathbb{E}X^k}{dp}.$$

Therefore it follows that the leading term of $\mathbb{E}X^{k+1}$ is

$$np\prod_{i=0}^{k-1}(n-i)p^k - kp^2\prod_{i=0}^{k-1}(n-i)p^{k-1} = \prod_{i=0}^{k}(n-i)p^{k+1}.$$

This proves the induction step and the lemma.