Moment-Based Learning of Mixture Distributions

SPUR Final Paper, Summer 2016

Kavish Gandhi and Yonah Borns-Weil
Mentor: Amelia Perry
Project suggested and supervised by Professor Ankur Moitra

August 4, 2016

Abstract

We study the problem of learning the parameters of a mixture of members of a given distribution family. To do this, we apply the method of moments, dating to Pearson in the late 1800's: we directly solve for the parameters in terms of estimated sample moments. We prove upper and lower bounds on the number of moments that uniquely determine mixtures for various distribution families. In particular, we show that $2k-1$ moments are necessary and sufficient to determine a large class of mixtures of k one-parameter distributions, including Poissons and exponentials, and develop an efficient algorithm to learn the parameters of these mixtures. Additionally, using natural exponential families as motivation, we ask which sets of $2k-1$ moments of a finite distribution uniquely determine its parameters, and show that this question can be reduced to asking about the zero sets of certain Schur polynomials. We also show that $4k-2$ moments are necessary for determining a mixture of k Gaussians, matching a known upper bound shown by Moitra and Valiant, and with this we improve the existing lower bound on the sample complexity of learning such a mixture. We prove similarly that $4k-2$ moments are necessary and sufficient to learn a mixture of k uniform distributions, and conjecture that a similar result holds for general two-parameter distributions whose moments satisfy certain polynomial dependence conditions on the parameters. Finally, for a general family of Gaussian-like distributions of the form $p(x)e^{q(x)}$, we derive a bound on the number of moments necessary to uniquely determine a mixture that is exponential in k.

1. Introduction

Background. A rich area of modern machine learning is the learning of mixture models, with applications throughout the scientific disciplines, including biology [1], finance [2], and physics [3, 4], among many others. The first results in the field came over 100 years ago in 1894, when biostatistician Karl Pearson studied the measurements of what he assumed was a single crab species and discovered an as-of-yet unseen distribution. His insight was that he was in fact observing *two* species, with each having measurements following the standard Gaussian distribution. This motivated defining a *(two) Gaussian Mixture Model* as a distribution F with density function $p_1 f_1 + p_2 f_2$, where $p_1 + p_2 = 1$ and f_1, f_2 are Gaussian density functions. To determine the parameters of the mixture from samples (and hence the data for each crab species), Pearson [5] invented the *method of moments*, in which he took the first five moments of a large number of samples drawn from this distribution, and solved by hand the resulting polynomial system for the means, variances, and probabilities. Getting finitely many solutions, he then chose the one whose sixth moment most closely matched the sixth sample moment. His results both demonstrated the effectiveness of learning a mixture from its moments and the utility of learning the parameters in particular, or *parameter learning*, rather than a distribution that is simply close in total variation distance, another rich area of study.

The problem of learning the parameters of a mixture of an arbitrary number of Gaussians has had a long history, where a mixture of k Gaussians is defined as the distribution with density function $\sum_{i=1}^{k} p_i f_i$, where $f_1, \ldots, f_k$ are the component Gaussian densities and $p_1, \ldots, p_k$ are mixing weights satisfying $\sum_{i=1}^{k} p_i = 1$. Many approaches to the problem since Pearson have involved clustering, starting with Dasgupta [6] in 1999, who used clustering to rigorously define a polynomial-time algorithm for separating d-dimensional Gaussians given a $\widetilde{\Omega}(\sqrt{d})$ separation between their means. The separation bound was gradually shrunk by later authors, including Arora and Kannan [7], Dasgupta and Schulman [8] and Vempala and Wang [9].

While these clustering results gave efficient algorithms for separated mixtures of Gaussians, they could not deal with even the simplest mixtures with lack of mean separation and non-negligible overlap of the components. Pearson's moment-based approach, however, proved fruitful for these types of mixtures, and in general for arbitrary Gaussian mixtures. In 2010, Moitra, Valiant, and Kalai [10] used a robust and rigorous version of Pearson's method of moments to give an algorithm that learns mixtures of two Gaussians to within parameter distance ϵ in time polynomial in ϵ, and in 2015 Hardt and Price [11] found an optimal algorithm for the two Gaussian mixture problem with respect to sample complexity, using an approach directly based on Pearson's polynomial. Moitra and Valiant [12], later in 2010, extended their result to mixtures of k Gaussians.

Simultaneously, Belkin and Sinha [13] generalized the moment method to any mixture of *polynomial families*, or families of distributions with moments polynomial in the parameters. Their proof was non-constructive, using the Hilbert basis theorem to show that finitely

many moments determine such mixtures, demonstrating that the method of moments is indeed a general, powerful strategy for learning mixtures of any polynomial family. However, the bound on a sufficient number of moments was simply shown to exist, and no effective bound was presented; the focus of this paper is on this question. Our goal is to establish effective bounds on the number of sufficient moments for particular families of distributions, independent of the parameters of the distribution.

Our results. Our work on moment-matching begins with the simplest case of *one-parameter families*, which are polynomial families determined by only a single parameter. The simplest one parameter distributions are the point masses, for which the corresponding mixtures are finite distributions; in Section 2, we review the classical theory of Gaussian quadrature, which shows that exactly $2k-1$ moments are required to uniquely determine such a finite distribution. Then, in Section 3, we study one-parameter distributions where the jth moment is a degree j polynomial in the parameter, and show that the problem of learning the parameters of a mixture of these distributions can be reduced to determining the finite distribution on the same parameters, for which we know $2k-1$ moments are sufficient. From these results, in Section 4 we develop an efficient, robust algorithm for learning mixtures of one-parameter subexponential distributions with this polynomial dependence on the moment, our first main result, Algorithm 4.1.

In Section 5, we review the theory of exponential families and natural exponential families, which have a particularly simple form for their density functions and comprise a large class of one-parameter families. By taking the mean as the parameter, these are automatically polynomial families, and the condition on the degree of the moments roughly corresponds to having the variance as a function of the mean $V(\mu)$ be quadratic, which gives the *natural exponential families with quadratic variance functions (NEF-QVF's)*, studied in [14] and which include many common distribution families. Using the method detailed in Algorithm 4.1, we therefore can efficiently find the parameters of almost any NEF-QVF mixture.

This discussion motivates us to look at NEF's with higher-degree polynomial variance functions, which reduces to the problem of determining finite distributions by certain linear combinations of their moments. In Section 6, we consider the simplified problem of determining for which $\{a_1, \dots a_{2k-1}\}$ the moments $\{M_{a_1}, \dots M_{a_{2k-1}}\}$ determine a finite distribution on k points. The main result of this section is that the moments of two distinct finite distributions differ whenever a certain Schur polynomial in the $2k$ points is nonzero, which implies the following result.

Theorem 6.6, Restated. Let $S = \{0, a_1, a_2, \dots, a_{2k-1}\}$ where $0 < a_1 < a_2 < \cdots < a_{2k-1} \in \mathbb{Z}$, and let $\lambda = (a_{2k-1}-(2k-1), a_{2k-2}-(2k-2), \dots, a_1-1, 0)$. If $s_\lambda(z_1, \dots, z_{2k})$ is nonzero for all real $z_1, \dots, z_{2k}$, then the moments $M_{a_1}, \dots M_{a_{2k-1}}$ uniquely determine a finite distribution on k points.

Thus, we can reduce one direction of this question to finding which Schur polynomials have only trivial zeros, and present some known partial results to that end, as well as the broader conjecture that this occurs whenever all parts in the partition are even.

In Section 7, we then return to the original problem of Gaussians, and place it in context of two-parameter distribution families in general. We show that $4k-2$ moments are sometimes necessary to determine a mixture of k Gaussians, matching the result in [10] that it is always sufficient.

Proposition 7.1. There exist two mixtures of k Gaussians, F, F', such that $F \neq F'$ and $M_j(F) = M_j(F')$ for $1 \leq j \leq 4k-3$.

The example for necessity, however, is a special case where all means are equal, so we conjecture that in the case of distinct means, only $3k$ moments are necessary. Along the lines of the sufficiency proof in [12], we also show that $4k-2$ moments are necessary and sufficient to uniquely determine a mixture of k uniform distributions, and conjecture that indeed $4k-2$ moments are sufficient to determine any mixture of a two-parameter family with moments that are degree at most k in each of the parameters. This conjecture is supported by some numerical evidence in the case of Gamma and Laplace distributions.

Finally, in Section 8, we consider the problem of learning a mixture of general families of distributions with an arbitrary number of parameters, in the hope of getting some concrete upper bound on the number of moments required to determine these mixtures, as opposed to the ineffective bound in [13]. We arrive at a result for a specific class of distribution families $\mathcal{D}$ of the form $p(x)e^{q(x)}$, when $p(x)$ and $q(x)$ are polynomials with degree independent of the parameters.

Proposition 8.2. Consider some $F \in \mathcal{D}$ with $\deg(p) = d_1 - 1$, $\deg(q) = d_2 > 0$. Then $O(d_1 d_2^{2k})$ moments suffice to uniquely determine a mixture of k F-distributions.

This leads us to our last conjecture, which would make explicit the ideas presented by Belkin and Sinha in [13] and give some general bound on the number of moments needed to match any mixture of members of a (not necessarily natural) exponential family, a class which includes nearly all named distribution families.

Conjecture 8.3. Let F be an exponential family. Then there exists a function $f(k)$ independent of the parameters such that at most $f(k)$ moments suffice to uniquely determine a mixture of k F-distributions.

2. Finite Distributions

In this section, we summarize the classical theory of moment matching for finite (i.e. finitely supported) distributions. In particular, we recall in Proposition 2.1 and 2.3 that $2k-1$ moments are both sufficient and necessary to uniquely determine a finite distribution on k distinct points, and outline an algorithm, whose main idea is given by Claim 2.2, that can exactly derive these points and weights given the first $2k-1$ moments. All of these results are well-known, although the construction in Proposition 2.3 is our own, but are stated for completeness.

Proposition 2.1. *If F and F' are finite distributions on k distinct points, and $M_j(F) = M_j(F')$ for $1 \leq j \leq 2k-1$, then $F = F'$.*

Proof. Let F have points $x_1, x_2, \ldots, x_k$ with weights $p_1, p_2, \ldots, p_k$, and F' have points $y_1, y_2, \ldots, y_k$ with weights $q_1, q_2, \ldots, q_k$. Suppose first that the distributions differ on at least one point; without loss of generality $y_k \neq x_i$ for all i. We claim that, in this case, there exists some j such that $1 \leq j \leq 2k-1$ and $M_j(F) \neq M_j(F')$.

Define the polynomial $P(x) = (x-x_1)\cdots(x-x_k)(y-y_1)\cdots(y-y_{k-1}) = \sum_{j=0}^{2k-1} a_j x^j$. Note that $P(x)$ evaluates to 0 at all points of the finite distributions except for y_k. Thus, we have that $\sum_{i=1}^{k}(p_i P(x_i) - q_i P(y_i)) = -q_k P(y_k) \neq 0$. But notice that we can also write this as

$$\begin{aligned}\sum_{i=1}^{k}(p_i P(x_i) - q_i P(y_i)) &= \sum_{j=1}^{2k-1} a_j \sum_{i=1}^{k}\left(p_i x_i^j - q_i y_i^j\right) \\ &= \sum_{j=1}^{2k-1} a_j (M_j(X) - M_j(Y)).\end{aligned}$$

For this to be nonzero, one of the moments must differ, as desired, so we have that if all of the points are not identical between distributions, the first $2k-1$ moments cannot be.

Now, consider the case in which the distributions share all of the same points; without loss of generality let $x_i = y_i$ for all $1 \leq i \leq k$. Letting $d_i = p_i - q_i$, and assuming for contradiction that all of the moments are equal but $F \neq F'$, we have that

$$\sum_{i=1}^{k} d_i x_i^j = 0$$

for all $j \leq k-1$ (in particular, this holds for $j \leq 2k-1$, but all we need is $j \leq k-1$). But this implies that $d_i = 0$ for all i by multiplying each side by the inverse of the Vandermonde matrix of $x_1, x_2, \ldots, x_k$, which exists because the x_i are distinct, giving that $p_i = q_i$ for all i, a contradiction of the fact that $F \neq F'$. Thus, we are done. □

Based on $2k-1$ moments uniquely determining a finite distribution, we conceivably could learn a finite distribution by evaluating its first $2k-1$ moments and algorithmically solving the polynomial system given by these moments to derive its parameters. Note that this is certainly not the most efficient way to learn a finite distribution, but will become useful in Section 3 when we talk about learning general one-parameter distributions.

Actually solving this polynomial system in the parameters given by the moments, which determines the underlying finite distribution, is implicit in the following well-known result from Gaussian quadrature, which generalizes Proposition 2.1.

Claim 2.2 (Gaussian Quadrature [15])**.** *Given moments $\mu_j = \int_a^b x^k dF(x)$ for $1 \leq j \leq 2k-1$ of a distribution F supported on $[a,b]$, there exist distinct $(x_1, \ldots, x_k)$, $(w_1, \ldots, w_k)$ satisfying $a \leq x_i \leq b$, $0 \leq w_i \leq 1$ for all i such that $\sum_{i=1}^{k} w_i x_i^j = \mu_j$ for all j and $\sum_{i=1}^{k} w_i = 1$.*

In particular, these $x_1, \ldots, x_k$ can be derived by defining the series of orthogonal polynomials $\{q_i\}_0^\infty$ defined by the recurrence

$$q_0(x) = (b-a)^{-1/2};\ q_{-1}(x) = 0;\ \beta_j q_j(x) = (x - \alpha_j)q_{j-1}(x) - \beta_{j-1}q_{j-2}(x),$$

where $\alpha_j = \int_a^b x q_{j-1}(x)^2 dF(x)$ and $\beta_j = \int_a^b x q_j(x) q_{j-1}(x) dF(x)$, known constants which can be evaluated in terms of the moments since the polynomials integrated have degree $\leq 2k-1$. The $x_1, \ldots, x_k$, then, are just the roots of q_{2k-1}. From here, to find the weights, we solve for w satisfying $Vw = M$, where

$$V = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_k \\ x_1^2 & x_2^2 & \cdots & x_k^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{k-1} & x_2^{k-1} & \cdots & x_k^{k-1}, \end{bmatrix}$$

is the Vandermonde matrix on $x_1, \ldots, x_k$, and M is the vector containing the first $k < 2k-1$ moments, an equivalent system to our moment polynomials. Since we are given that the x_i are distinct, $\det(V) = \prod_{1 \leq i < j \leq n} (x_j - x_i)$ is nonzero, which means that we can solve for w simply as $V^{-1}M$.

Using this, we can learn a finite distribution F on k points given $2k-1$ moments as follows: first, let $a = -M_2(F), b = M_2(F)$, so that our range of integration indeed contains all of the x_i. Using the above recurrence, we can derive the sequence of orthogonal polynomials to F; our points $x_1, \ldots, x_k$ are then the roots of the $(2k-1)$st orthogonal polynomial, and our weights can be found by evaluating $V^{-1}M$. Since the first $2k-1$ moments of the finite distribution defined with these points and weights are exactly those of the finite distribution we are trying to learn, by Proposition 2.1, our derived finite distribution is correct.

The algorithm above allows us to precisely determine the parameters of a finite distribution, given the first $2k-1$ moments. We will use this as a baseline to develop a robust, general algorithm in Section 4 for learning mixtures of distributions belonging to one-parameter families. In particular, the described algorithm only works when the moments are known exactly; for one-parameter families, we will develop a robust algorithm that estimates the parameters up to some error ϵ when the moments are known up to a specified polynomial in ϵ and a parameter of the distribution in question. Notice, however, that we cannot do better with exact moment-matching for finite distributions because of the following well-known result, whose proof is deferred to the appendix.

Proposition 2.3. *There exist two finite distributions F, F', such that $F \neq F'$ and $M_j(F) = M_j(F')$ for $1 \leq j \leq 2k-2$.*

3. One Parameter Families

In this section, we use Proposition 2.1 to show that $2k-1$ moments actually uniquely determine a mixture of k components for a large class of one-parameter distributions, because we can relate their moments to those of an associated finite distribution.

In particular, let $\mathcal{C}$ be a class of one-parameter distribution families whose jth moment is a degree j polynomial in the parameter for all $j \geq 0$.

Definition 3.1. Given parameters $\lambda_1, \lambda_2, \ldots, \lambda_k$ and mixing probabilities $p_1, p_2, \ldots, p_k$, a $(\mathcal{C}, k)$-*mixture* is the distribution given by sampling from X_i with probability p_i for all $1 \leq i \leq k$, where the X_i all belong to some $C \in \mathcal{C}$.

Given a $(\mathcal{C}, k)$-mixture X, we will denote by $\mathrm{IF}(x)$ the finite distribution specified by atoms at $\lambda_1, \ldots, \lambda_k$ and weights $p_1, \ldots, p_k$.

Proposition 3.2. *Let X, X' be $(\mathcal{C}, k)$-mixtures with components belonging to the same $C \in \mathcal{C}$. Then $M_j(X) = M_j(X')$ for $0 \leq j \leq 2k-1$ if and only if $M_j(\mathrm{IF}(X)) = M_j(\mathrm{IF}(X'))$ for all $1 \leq j \leq 2k-1$.*

Proof. To prove this, we will show that, given a $(\mathcal{C}, k)$-mixtures X, we can solve for moments of X given the moments of $\mathrm{IF}(X)$, and vice versa. This implies the desired result.

If the components of X are $X_1, \ldots, X_k$, then $M_j(X) = \sum_{i=1}^{k} p_i M_j(X_i)$. Since X_i is a C-distribution, $M_j(X_i) = \sum_{m=0}^{j} \alpha_m \lambda_i^m$ with $\alpha_j \neq 0$, so we have that

$$M_j(X) = \sum_{i=1}^{k} p_i \left(\sum_{m=0}^{j} \alpha_m \lambda_i^m \right) = \sum_{m=0}^{j} \alpha_m \left(\sum_{i=1}^{k} p_i \lambda_i^m \right) = \sum_{m=0}^{j} \alpha_m M_m(\mathrm{IF}(X)),$$

so the moments of X can indeed be determined from those of $\mathrm{IF}(X)$.

Oppositely, we show by induction that $M_j(\mathrm{IF}(X))$ can be expressed as a linear combination of $M_i(X)$ for $0 \leq i \leq j$. The base case is clearly true; now assume it is true for all $i < j$ that $M_i(\mathrm{IF}(X)) = \sum_{l=0}^{i} \alpha_l M_l(X)$. Now, notice that, as shown previously, $M_j(X) = \sum_{m=0}^{j} \alpha_m M_m(\mathrm{IF}(X))$ for constants α_m and $\alpha_j \neq 0$; moving all of the $M_m(\mathrm{IF}(X))$ terms satisfying $m < j$ to the right-hand side and expressing them as linear combinations of $M_i(X)$ satisfying $0 \leq i \leq m$, possible by our inductive hypothesis, we are done. □

Note that, by Proposition 2.1, $M_j(F) = M_j(F')$ for all $1 \leq j \leq 2k-1$ if and only if $F = F'$, provided the parameters are distinct, which implies that $X = X'$ because, since $F = F'$, they share the same parameters and weights. From this, we have the following:

Corollary 3.3. *Given two $(\mathcal{C}, k)$-mixtures X, X' of components with distinct parameters, $X = X'$ if and only if $M_j(X) = M_j(X')$ for $0 \leq j \leq 2k-1$.*

4. A Robust Algorithm for learning $(\mathcal{C}, k)$-mixtures

In this section, based on the results from Section 3, we present an robust, efficient moment-matching algorithm to learn a $(\mathcal{C}, k)$-mixture. Corollary 3.3 implies that, once we learn the

$2k-1$ moments of a $(\mathcal{C},k)$-mixture, this mixture is uniquely determined. Our first result makes this result robust by giving us an estimate of how well we need to learn the moments to determine the mixture up to $\pm\epsilon$ in the parameters.

We first show that estimating the moments of a $(\mathcal{C},k)$ mixture well allows to also estimate the moments of the corresponding finite distribution well. Note that throughout this section we will assume k is held fixed, so will treat it as a constant.

Proposition 4.1. *Given an estimate $\widehat{M}_j(X)$, $1 \le j \le 2k-1$, of the first $2k-1$ moments of X satisfying $|\widehat{M}_j(X) - M_j(X)| \le \beta$, Algorithm 4.1 outputs an estimate of $\widehat{M}_j(\mathrm{IF}(X))$ satisfying $|\widehat{M}_j(\mathrm{IF}(X)) - M_j(\mathrm{IF}(X))| < O(\beta)$.*

Proof. Following exactly the forward implication of Proposition 3.2, we can derive an estimate $\widehat{M}_j(\mathrm{IF}(X))$ with the following recursive formulation:

$$\widehat{M}_j(\mathrm{IF}(X)) = \begin{cases} 1 & \text{if } j = 0 \\ \frac{\widehat{M}_j(X) - \sum_{i=0}^{j-1} \alpha_i \widehat{M}_i(\mathrm{IF}(X))}{\alpha_j} & \text{otherwise} \end{cases}.$$

We will prove the desired robustness statement by induction. The base case is trivial. Now, assume, for $0 \le i \le j-1$, we know all of the $\widehat{M}_i(\mathrm{IF}(X))$ are estimates up to $c_i\beta$ for some constants c_i. By assumption, we know $\widehat{M}_j(X)$ up to β, so we have that, using that the α_j are constants (since we are holding k fixed),

$$\begin{aligned} |\widehat{M}_j(\mathrm{IF}(X)) - M_j(\mathrm{IF}(X)| &= \frac{1}{\alpha_j}\left|\left(\widehat{M}_j(X) - \sum_{i=0}^{j-1}\alpha_i\widehat{M}_i(\mathrm{IF}(X))\right) - \left(M_j(X) - \sum_{i=0}^{j-1}\alpha_i M_i(\mathrm{IF}(X))\right)\right| \\ &\le \frac{1}{\alpha_j}\left(\left|\widehat{M}_j(X) - M_j(X)\right| + \sum_{i=0}^{j-1}\alpha_i\left(\left|\widehat{M}_i(\mathrm{IF}(X)) - M_i(\mathrm{IF}(X))\right|\right)\right) \\ &\le \frac{1}{\alpha_j}\left(\beta + \sum_{i=0}^{j-1}\alpha_i c_i \beta\right) \\ &= O(\beta). \end{aligned}$$

□

Now, similar to as defined by Moitra and Valiant [12], we will show that pairs of finite distributions satisfying certain separation conditions must have a moment that differs by a certain amount. This will imply that their corresponding $(\mathcal{C},k)$ mixtures must also have a similar moment, as desired.

Definition 4.2. We say that a finite distribution F is ϵ-separated if the following two conditions hold:

1. $|x_i - x_j| \ge \epsilon$ for all $i \ne j$.
2. $p_i \in [\epsilon, 1]$.

Definition 4.3. A pair of finite distributions F, F' is said to be (ϵ_1, ϵ_2)-standard if the following conditions hold:

1. F and F' are ϵ_1-separated.
2. $|x_i, x_i'| \leq \frac{1}{\epsilon_1}$.
3. $\epsilon_2 \leq \min_\pi \left(\sum_{i=1}^{k} |x_i - x'_{\pi(i)}| + |p_i - p'_{\pi(i)}| \right)$,

where the above minimization is taken over all permutations $\pi : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, k\}$.

Theorem 4.4. *Given a (ϵ_1, ϵ_2)-standard pair of finite distributions F, F' with $\epsilon_1 \gg \epsilon_2$, there exists some i satisfying $1 \leq i \leq 2k-1$ such that*

$$|M_i(F) - M_i(F')| > \Omega(\epsilon_1^{4k-2}\epsilon_2)$$

The proof of this theorem is deferred to Appendix A, but the proof idea is a robust version of that of Proposition 2.1; we construct a polynomial that goes through all but one or two of the points of F, F', and show that $\int_x |p(x)f(x)|$ is at least $\Omega(\epsilon_1\epsilon_2^{2k-1})$, which will in turn give two moments that differ by $\Omega(\epsilon_1\epsilon_2^{4k-2})$, from the upper bound on the x_i.

The final step in developing an algorithm to learn mixtures of distributions in $\mathcal{C}$ is to show that we can actually estimate the moments of a $(\mathcal{C}, k)$ mixture with a minimal number of samples. This may not always be true, but it turns out we can estimate the moments with probability $1 - \delta$ given $O\left(\epsilon^{-2} \log\left(\frac{1}{\delta}\right)\right)$ samples, as proven in Proposition 4.6, when the distribution is subexponential, defined as follows.

Definition 4.5. A distribution F is *subexponential* with parameter λ if the kth raw moment satisfies $M_k \leq k^k\lambda^k$.

There are many equivalent definitions of subexponential distributions and the associated parameter, which serve different purposes. Subexponential distributions intuitively are meant to be those distributions with an exponential or lighter tail, and encompass many of the common distributions that we study.

Note that a mixture of subexponential distributions is clearly also subexponential, where the associated subexponential parameter is at most the maximum among its components.

Proposition 4.6. *If a distribution $F \in \mathcal{C}$ is subexponential with parameter λ, then given $t = O(1)$ there exists an algorithm taking $n = O\left(\epsilon^{-2} \log\left(\frac{1}{\delta}\right)\right)$ samples that with probability $1 - \delta$ learns the ith moment, for $i \leq t$ to within $\epsilon\lambda^i$.*

The proof of this result is deferred to Appendix A, since it is very similar to that of [11], but note that this result implies the following corollary, where we replace λ with the standard deviation σ, since $\sigma^2 \geq \sum_{i=1}^{k} p_i\sigma_i^2$ from the law of total variance.

Corollary 4.7. *Given a one-parameter distribution $C \in \mathcal{C}$ with parameter λ whose variance σ' is quadratic in λ and $t = O(1)$, then there exists an algorithm taking $n = O\left(\epsilon^{-2} \log\left(\frac{1}{\delta}\right)\right)$ samples that with probability $1 - \delta$ learns the ith moment of a mixture F of variance $\sigma > 1$ of k component C-distributions for $i \leq t$ to within $\epsilon\sigma^i$.*

Using all of our results, we can state a robust algorithm, Algorithm 4.1, that learns the parameters of a $(\mathcal{C}, k)$ mixture up to ϵ, given that it is subexponential.

Theorem 4.8. *Let X be a $(\mathcal{C}, k)$, ϵ_1-separated mixture of components $X_1, \ldots, X_k$ belonging to some subexponential $C \in \mathcal{C}$ with distinct parameters satisfying $\lambda_i \leq \frac{1}{\epsilon_1}$ for $\epsilon_1 < 1$. Then, given any $\epsilon_2 > 0$ satisfying $\epsilon_2 \ll \epsilon_1$ and a subexponential parameter $\gamma \geq 1$ for C, Algorithm 4.1 recovers the λ_i, w_i to $\pm\epsilon_2$ with probability $1 - \delta$.*

Proof. By Proposition 4.6, given n samples, steps 1 and 2 estimate the moments $\widehat{M_j}(X)$, for $1 \leq j \leq 2k - 1$, up to

$$(a\epsilon_2^{-2}\epsilon_1^{4-8k}\gamma^{4k-2})^{-0.5}\gamma^j = a^{-0.5}\epsilon_2\epsilon_1^{4k-2}\gamma^{1-2k}\gamma^j < a^{-0.5}\epsilon_2\epsilon_1^{4k-2}.$$

By Proposition 4.1, this means that we have estimates $\widehat{M_j}(\mathrm{IF}(X))$ of the moments of the associated finite distribution that are also accurate up to $\pm b\epsilon_2\epsilon_1^{4k-2}$, for b a constant. Choosing a suitably given the family C and k, Theorem 4.4 gives us that the pair of $F = \mathrm{IF}(X)$ and the finite distribution F' on $\widehat{M_j}(\mathrm{IF}(X))$ are not (ϵ_1, ϵ_2)-standard, since their moments are too close, which gives that there exists a permutation π such that $|\lambda_i - \hat{\lambda}_{\pi(i)}| < \epsilon_2$ and $|w_i - \hat{w}_{\pi(i)}| < \epsilon_2$ for all $1 \leq i \leq k$, as desired.

The remaining steps of the algorithm concern actually deriving F', the finite distribution on $\widehat{M_j}(\mathrm{IF}(X))$, the correctness of which was given by Proposition 2.1 and Claim 2.2 and the steps of which outlined in Section 2. It is worth mentioning that, because our estimates $\widehat{\lambda_i}$ are correct up to $\pm\epsilon_2$ and $|\lambda_i - \lambda_j| \geq \epsilon_1$ by assumption for $i \neq j$, $\hat{V}$ is still invertible. Furthermore, a and b are chosen such that $|a, b| \geq M_2(\mathrm{IF}(X))$, which implies that the values of λ_i lie in the integration range, as required. □

Algorithm 4.1: Efficiently learning $(\mathcal{C}, k)$ mixtures

Input : An accuracy parameter ϵ_2, error δ, and a $(\mathcal{C}, k)$ mixture, X, of components $X_1, \ldots, X_k$ belonging to some $C \in \mathcal{C}$, to which we are provided sample access, which is ϵ_1-separated and has parameters satisfying $|\lambda_i| \leq \epsilon_1$ for $\epsilon_1 \gg \epsilon_2$, and parameter γ, the subexponential parameter.

Output: Estimates of the $2k$ parameters of the components of X, (λ_1, ..., λ_k), $(w_1, \ldots, w_k)$ to within $\pm\epsilon_2$ with probability $1 - \delta$.

1 Draw $n = a\left(\epsilon_2^{-2}\epsilon_1^{4-8k}\gamma^{4k-2}\log(1/\delta)\right)$ samples $x_1, \ldots, x_n$ from X, and split them into $\log(1/\delta)$ groups G_i of size $s = a\epsilon_2^{-2}\epsilon_1^{4-8k}\gamma^{4k-2}$, where a is a constant depending on the family C and k.

2 For $1 \leq j \leq 2k-1$, compute the empirical moments $\widehat{M_j}(G_i) = \frac{1}{s}\sum_{i=1}^{s} x_i^j$ of each group G_i, and estimate the moment $\widehat{M_j}(X)$ as the median of these empirical moments.

3 Convert these estimated moments into empirical moments of the finite distribution $\mathrm{IF}(X)$ on the parameters from the following recursive formulation:

$$\widehat{M_j}(\mathrm{IF}(X)) = \begin{cases} 1 & \text{if } j = 0 \\ \frac{\widehat{M_j}(X) - \sum_{i=0}^{j-1} \alpha_i \widehat{M_i}(\mathrm{IF}(X))}{\alpha_j} & \text{otherwise} \end{cases}.$$

4 Given that $M_2(Y) = \alpha_2\kappa^2 + \alpha_1\kappa + \alpha_0, M_1(Y) = \kappa$ for a distribution $Y \in C$ with parameter κ, let $b = -a = \frac{1}{\alpha_2}(\widehat{M_2}(X) - k\alpha_0 - \alpha_1\widehat{M_1}(X) + 2k\epsilon)$. This ensures that all of the λ_i are in the correct range for the following step.

5 Derive the first k orthogonal polynomials $\{q_i\}_0^k$ to this distribution defined by the recurrence

$$q_0(x) = (b-a)^{-1/2};\ q_{-1}(x) = 0;\ \beta_j q_j(x) = (x - \alpha_j)q_{j-1}(x) - \beta_{j-1}q_{j-2}(x),$$

where $\alpha_j = \int_a^b x q_{j-1}(x)^2 d\mathrm{IF}(x)$ and $\beta_j = \int_a^b x q_j(x) q_{j-1}(x) d\mathrm{IF}(x)$.

6 Compute the roots of q_{2k}; these are the estimates of the parameters $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_k$.

7 Letting

$$\widehat{V} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \widehat{\lambda_1} & \widehat{\lambda_2} & \cdots & \widehat{\lambda_k} \\ \widehat{\lambda_1}^2 & \widehat{\lambda_2}^2 & \cdots & \widehat{\lambda_k}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\lambda_1}^{k-1} & \widehat{\lambda_2}^{k-1} & \cdots & \widehat{\lambda_k}^{k-1} \end{bmatrix}, \quad \widehat{M} = \begin{bmatrix} \widehat{M_0}(F) \\ \widehat{M_1}(F) \\ \vdots \\ \widehat{M_{k-1}}(F) \end{bmatrix},$$

let $\widehat{w} = \widehat{V}^{-1}\widehat{M}$. Our estimate of the weight $\widehat{w_i}$ is the ith coordinate of this vector.

5. Natural Exponential Families

Our result on one-parameter families that satisfy a particular polynomial condition on their moments may seem unmotivated, since the condition does not obviously correspond to a particular family of distributions. However, at least a subset of these one-parameter families belong to a more general class of families of distributions parameterized by some parameter vector, known as exponential families. In this section, we describe a special case of these, natural exponential families, and use this case to give examples of distribution families in $\mathcal{C}$.

Definition 5.1. Consider a family of d-dimensional distributions parameterized by $\theta \in \mathbb{R}^s$, with probability density functions f_{X_θ}. We call this an *exponential family* if we can find functions $h : \mathbb{R}^d \to \mathbb{R}, \eta : \mathbb{R}^s \to \mathbb{R}^s, T : \mathbb{R}^d \to \mathbb{R}^s, A : h : \mathbb{R}^s \to \mathbb{R}$ such that

$$f_{X_\theta}(x) = h(x)e^{\eta(\theta)\cdot T(x) - A(\theta)}.$$

We call η the *natural parameter*, T the *sufficient statistic*, and A the *log-partition function*.

This is an enormous class of parametric families of distributions, including many common distributions such as Gaussians, Poisson distributions, and Gamma distributions, so we can restrict a bit and still get a quite rich collection.

Definition 5.2. An exponential family is called a *natural exponential family* if T and η are both the identity function.

Specifically, a natural exponential family has a density of the form

$$f_{X_\theta}(x) = h(x)e^{\theta x - A(\theta)}.$$

In this case, θ can be thought of as the parameter, though it is important to keep in mind that it may not exactly correspond to the standard definition of the parameter for that particular distribution.

Natural exponential families are especially motivated because they are described completely by their mean or variance, as shown by the following facts, of which one can find a fuller discussion in [14].

Fact 5.3. *If A is the log-partition function in a natural exponential family, then the rth cumulant is given by $A^{(r)}(\theta)$.*

Notice that this gives that $A''(\theta) > 0$, so A' is injective and hence invertible. From this, we can define the *variance function* $V(\mu)$ as a function of the mean $\mu = A'(\theta)$, which implies the following fact.

Fact 5.4. *The variance function V, together with a domain, determines the natural exponential family.*

It is also true, for a natural exponential family, that writing the variance function allow for a simple recurrence formula for cumulants of a member of a natural exponential family.

Fact 5.5.

$$C_{j+1}(\mu) = V(\mu)C_j'(\mu)$$

where $C_k(\mu)$ is the kth cumulant of the distribution with mean μ.

Since we wish to work with families in $\mathcal{C}$, the recurrence above motivates us to work with *natural exponential families with quadratic variance functions (NEF-QVF's)*, which are natural exponential families where the variance is a quadratic polynomial in the mean. By transforming and considering cases, Morris [14] showed that up to division, convolution, and affine transformations, there are just 6 NEF-QVF's:

1. The Gaussian distribution with fixed variance σ^2,
2. The Poisson distribution,
3. The Gamma distribution with fixed shape parameter α,
4. The Binomial distribution with fixed number of trials n,
5. The Negative Binomial distribution with fixed failure number r,
6. The Generalized Hyperbolic Secant distribution with fixed r.

For all of these distributions except for the binomial distribution, the jth moments are precisely jth degree polynomials in the mean (which can be taken as the defining parameter), giving that they can be learned efficiently using Algorithm 4.1. More details on each of the NEF-QVF's are given in Appendix B.

Based on this connection between natural exponential families with quadratic variance and distribution families in $\mathcal{C}$, we are then encouraged to consider, as a next step, mixtures of natural exponential families with higher-degree polynomial variance functions, which include the well-known Wald (Inverse Gaussian) distributions with cubic variance, which, when the scaling parameter is fixed to 1, has moments $M_1 = \mu$, $M_2 = \mu^3 + \mu^2$, $M_3 = \mu^5 + \mu^4 + \mu^3$, which satisfy the recurrence $M_{n+1} = (2n-1)\mu^2 M_n + \mu^2 M_{n-1}$. These are not $(\mathcal{C}, k)$-mixtures, so our results do not apply, but perhaps we can alter the same technique to deal with these cases as well. This is the subject of Section 6.

6. Learning a finite distribution with some $2k-1$ moments

Up to this point, when learning finite distributions and a mixture of one-parameter distributions, we assumed that we would be able to reduce the mixture to the problem of solving for a finite distribution given its first $2k-1$ moments, which we were able to solve robustly. However, as the inverse Gaussian shows us, we often encounter situations in which our reduction may in fact involve a set of $2k-1$ moments that are not the first $2k-1$ moments, or a linear combination thereof.

A pertinent starting question, then, is: which sets of $2k-1$ moments $\{M_{a_1}, M_{a_2}, \ldots, M_{a_{2k-1}}\}$ uniquely determine a finite distribution F on k points?

To approach this question, we generalize the approach outlined in the proof of Proposition 2.1, in which, given a pair of unequal F, F', we construct a polynomial $p(x)$ that has roots at all but one of the points of F, F', thus giving a nonzero integral $\int_x |p(x)(F - F')|$. In Proposition 2.1, we wanted this polynomial to be of degree $2k-1$, since this implied that one of the first $2k-1$ moments differs. In this case, we want our polynomial to have at most $2k-1$ nonzero coefficients, corresponding to $x^{a_1}, x^{a_2}, \ldots, x^{a_{2k-1}}$. We first show that it is possible to construct such a polynomial from the $p(x)$ given in Proposition 2.1.

Lemma 6.1. *Let $S = \{0, a_1, a_2, \ldots, a_{2k-1}\}$ where $0 < a_1 < a_2 < \cdots < a_{2k-1} \in \mathbb{Z}$. Given distinct finite distributions F, F' on points $x_1, \ldots, x_k$ and $y_1, \ldots, y_k$, there exists a nonzero polynomial $q(x) = \sum_{i \in S} c_i x^i$ such that $q(x)$ is divisible by $\prod_{i=1}^{k}(x - x_i) \prod_{i=1}^{k-1}(x - y_i)$.*

Proof. Let $D = \{d_1, \ldots, d_l\} = \{0, 1, \ldots, a_{2k-1}\} \setminus S$. Let

$$U(x) = \sum_{i=0}^{2k-1} u_i x^i = \prod_{i=1}^{k}(x - x_i) \prod_{i=1}^{k-1}(x - y_i),$$

and M be the $l \times (l+1)$ matrix such that $M_{ij} = u_{d_i - j}$. Because the numbers of rows of M is l, we have that $\text{rank}(M) \leq l$, which gives that $\text{nullity}(M) \geq 1$. This means that there must exist some nonzero vector $b = (b_0, b_1, \ldots, b_l)$ satisfying $Mb = 0$. Then letting $B(x) = \sum_{i=0}^{l} b_i x^i$ gives $q(x) = B(x)U(x)$ to satisfy the desired properties. □

It is useful to know that we can always construct such a polynomial; our goal, however, is to construct such a polynomial that has exactly $2k-1$ roots in P, since this gives us the following result.

Proposition 6.2. *Let $S = \{0, a_1, a_2, \ldots, a_{2k-1}\}$ where $0 < a_1 < a_2 < \cdots < a_{2k-1} \in \mathbb{Z}$. Given finite distributions F, F' on distinct points $P = \{x_1, \ldots, x_k, y_1, \ldots, y_k\}$ such that $F \neq F'$, there exists a nonzero polynomial $q(x) = \sum_{i \in S} c_i x^i$ such that $q(x)$ has shares as roots all but one of the elements of P. Then the set of moments $\{M_{a_1}, M_{a_2}, \ldots, M_{a_{2k-1}}\}$ uniquely determines a finite distribution on k points.*

Proof. This follows immediately from the fact that the integral $\int_x |q(x)(F - F')|$ is nonzero, implying that there must exist some $i \in S$ such that $M_i(F) \neq M_i(F')$, since these correspond to the nonzero coefficients of q. □

One particular instance in which this polynomial is easy to construct occurs when the a_i are uniformly spaced with an odd difference; in other words, $a_i = a_{i-1} + j$ for some odd $j \in \mathbb{Z}^+$.

Proposition 6.3. *Let $S = \{a_1, a_2, \ldots, a_{2k-1}\} = \{a, a + j, \ldots, a + (2k-2)j\}$ where j is an odd positive integer and $a > 0$. Then the set of moments $\{M_{a_1}, M_{a_2}, \ldots, M_{a_{2k-1}}\}$ uniquely determines a finite distribution on k points.*

Proof. Let F, F' be two finite distributions on points $x_1, \ldots, x_k, y_1, \ldots, y_k$, where $x_a \neq x_b, y_a \neq y_b$ for all $a \neq b$. First, assume that there exists some l such that $y_l \neq x_i$ for all $1 \leq i \leq k$. Without loss of generality, let $k = l$, and let y_k be nonzero, possible because, given the above, there also exists some x_j which is distinct from all of the y_i for $1 \leq i \leq k$; without loss of generality we assume that y_k is nonzero, since we know that $x_j \neq y_k$. With this established, apply Proposition 6.2 using the polynomial

$$q_1(X) = x^a \prod_{l=0}^{j-1} \left(\prod_{i=1}^{k} (x - x_i \zeta_j^l) \prod_{i=1}^{k-1} (x - y_i \zeta_j^l) \right)$$

where $\zeta_j = e^{i\frac{2\pi}{j}}$, possible because the only real roots of this are 0, $x_1, \ldots, x_k$, $y_1, \ldots, y_{k-1}$.

If there exists no such l, then without loss of generality let $x_i = y_i$ for all i. Because $F \neq F'$, there must exist some j such that $p_j \neq q_j$; without loss of generality, let $j = k$. Then, similarly define the polynomial

$$q_2(X) = x^a \prod_{l=0}^{j-1} \left(\prod_{i=1}^{k-1} (x - x_i \zeta_j^l) \prod_{i=1}^{k-1} (x - y_i \zeta_j^l) \right)$$

and apply Proposition 6.2, noting that the integral $\int_x |q_2(x)(F - F')|$ is still nonzero because of the difference in weights. □

When the a_i are uniformly spaced with an even difference, then it is easy to construct counterexamples in which two different mixtures share the same set of moments, since either the moments are all even or all odd. In particular, if all of the moments are even, then given a mixture F with component parameters x_i, p_i, the mixture F' with parameters $-x_i, p_i$ clearly has the same moments. Oppositely, if all of the moments are even, then, given a mixture F with component parameters $(x_1, -x_1, x_2, \ldots, x_{k-1}), (p_1, p_1, p_2, \ldots, p_{k-1})$, then the mixture F' with component parameters $(y_1, -y_1, x_2, \ldots, x_{k-1}), (p_1, p_1, p_2, \ldots, p_{k-1})$ has the same moments but is a distinct mixture if we pick $y_1 \neq x_1$.

In general, we can also approach this problem as follows: from Proposition 6.1, we know that it is always possible to extend a polynomial $g(x)$ with roots at all but one of the points of the two mixtures to another polynomial $g^\star(x)$ with nonzero coefficients only at the known moments. If the same cannot be done for $h(x) = \prod_{i=1}^{k}(x - x_i)(x - y_i)$, which has all of the points as roots, then $g^\star(x)$ cannot be divisible by $h(x)$, which means that $\int_x |g^\star(x)(F - F')|$ must be nonzero, as desired. Thus, one approach is to determine for which sets $\{a_1, a_2, \ldots, a_{2k-1}\}$ we can show that $h(x)$ cannot be extended.

It turns out that a sufficient condition involves the *Schur polynomials* $s_\lambda(x_1, x_2, \ldots, x_k, y_1, \ldots, y_k)$, where λ is the partition $(\lambda_1, \lambda_2, \cdots, \lambda_{2k-1}, 0)$, $\lambda_i = a_{2k-i} - (2k - i)$. These polynomials have a number of equivalent definitions; we will present the one most useful for our discussion.

Definition 6.4. A *generalized Vandermonde matrix* $V_\lambda(x_1, \ldots, x_k)$ on $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k)$

and coordinates $x_1, x_2, \ldots, x_k$ is defined as

$$\begin{bmatrix} x_1^{\lambda_1+k-1} & x_1^{\lambda_2+k-2} & \cdots & x_1^{\lambda_k} \\ x_2^{\lambda_1+k-1} & x_2^{\lambda_2+k-2} & \cdots & x_2^{\lambda_k} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^{\lambda_1+k-1} & x_k^{\lambda_2+k-2} & \cdots & x_k^{\lambda_k} \end{bmatrix}.$$

Definition 6.5. Given a partition $(\lambda_1, \lambda_2, \cdots, \lambda_n)$, the Schur polynomial s_λ is defined as

$$s_\lambda(x_1, x_2, \ldots, x_n) = \frac{\det(V_\lambda(x_1, \ldots, x_n))}{\det(V_{(0)}(x_1, \ldots, x_n))}.$$

Using this definition, we can prove the following.

Theorem 6.6. *Let $S = \{0, a_1, a_2, \ldots, a_{2k-1}\}$ where $0 < a_1 < a_2 < \cdots < a_{2k-1} \in \mathbb{Z}$, and let $\lambda = (a_{2k-1} - (2k-1), a_{2k-2} - (2k-2), \ldots, a_1 - 1, 0)$. Given two finite distributions F, F' on distinct points $x_1, \ldots, x_k, y_1, \ldots, y_k$, there exists a nonzero polynomial $q(x) = \sum_{i \in S} c_i x^i$ divisible by $\prod_{i=1}^k (x - x_i) \prod_{i=1}^k (x - y_i)$ if and only if $s_\lambda(x_1, \ldots, x_k, y_1, \ldots, y_k) = 0$.*

Proof. Consider $V_\lambda(x_1, \ldots, x_k, y_1, \ldots, y_k)$, where λ, as before, is the partition $(\lambda_1, \lambda_2, \cdots, \lambda_{2k})$, where $\lambda_i = a_{2k-i} - (2k - i)$ and $\lambda_{2k} = 0$. Note that, on distinct x_i, y_i, we have that $V_0(x_1, \ldots, x_k, y_1, \ldots, y_k)$ is nonzero, so $\det(V_\lambda(x_1, \ldots, x_k, y_1, \ldots, y_k))$ is zero if and only if $s_\lambda(x_1, \ldots, x_k, y_1, \ldots, y_k) = 0$. But this determinant being zero is equivalent to a nontrivial linear relationship among the rows of M, which gives a $q(x)$ with the desired nonzero coefficients with roots at $x_1, \ldots, x_k, y_1, \ldots, y_k$, as desired. Since all of these steps are reversible, both the forward and reverse implications hold. □

In a particularly simple case, we can consider $\lambda = (m, 0, 0, \ldots, 0)$ where there are $n \geq m - 1$ zeros in the partition. Then

$$s_\lambda = \sum_{0 \leq i_1 \cdots \leq i_m \leq n} \left(\prod_{j=1}^m x_{i_j} \right)$$

which is known as the *complete homogeneous symmetric polynomial* h_m. Hunter [16] proved the following:

Theorem 6.7. *The polynomial h_{2r} is positive except at zero.*

Corollary 6.8. *Let $S = \{1, 2, \ldots, 2k-2, 2j+1\}$, where $j \geq k$. Then the set of moments M_S uniquely determines a finite distribution on k points.*

Proof. This follows from Theorem 6.7, Theorem 6.6, and Proposition 6.2. □

Notice also that Proposition 6.3 implies that we can perhaps show that, given $\lambda = ((n-1)j, \ldots, j, 0)$ for j an even nonnegative integer, we have that $s_\lambda(z_1, \ldots, z_n) > 0$ for any choice of real $z_1, \ldots, z_n$. In fact, this is indeed the case; the complete proof is omitted for the sake of space, but it leads us to believe that the following conjecture is true.

Conjecture 6.9. *If λ is a partition with only even parts and last part 0, then $s_\lambda(x_1, \ldots, x_k, y_1, \ldots, y_k)$ is positive away from zero.*

7. Two-Parameter Mixtures

Our results to this point have concerned only one-parameter distributions; in this section, we consider how these techniques extend two-parameter distributions. In particular, Moitra and Valiant [12] previously showed that $4k-2$ moments suffice to uniquely determine a mixture of k Gaussians; we show that this is a necessary condition.

Proposition 7.1. *There exist two mixtures of k Gaussians, F, F', such that $F \neq F'$ and $M_j(F) = M_j(F')$ for $1 \leq j \leq 4k-3$.*

Proof. Let the means of F and F', respectively, be denoted $\mu_1, \mu_2, \ldots, \mu_k$ and $m_1, m_2, \ldots, m_k$, and the variances be denoted $\sigma_1^2, \ldots, \sigma_k^2$ and $\nu_1^2, \ldots, \nu_k^2$. Let $\mu_1 = \mu_2 = \cdots = \mu_k = m_1 = m_2 = \cdots = m_k = 0$, which causes the odd moments of F and F' to all be 0, and the even moments $M_{2j}(F), M_{2j}(F')$ to be $\sum_{i=1}^{k} p_i \sigma_i^{2j}$, $\sum_{i=1}^{k} p_i \nu_i^{2j}$; in particular, they represent the moments of finite distributions at points σ_i^2, ν_i^2. From Proposition 2.3, then, given σ_i^2, we can choose ν_i^2 (we can ensure positivity by taking our interval to be $[0, 1]$, which by Proposition 2.2 implies that our ν_i^2 fall in that range) to match the first $2k-2$ moments of these finite distributions, which we implies that we match up to the $(4k-4)$th even moment of F, F'. Since the $(4k-3)$rd moment and indeed all of the odd moments are matched by the fact that they all are zero, we have the desired result. □

Note that the above construction can actually match $4k-3$ moments for any mixture of k Gaussians in which the means are all equal. We can see this by the fact that the above mixtures having equal moments implies they have equal cumulants, and we can shift the first cumulant, the mean, without changing the later cumulants, implying the desired result.

However, this matching of many moments seems to be specific to the case in which many of the means are equal; we suspect that as the means are separated, the number of moments that suffice to determine the distribution becomes smaller:

Conjecture 7.2. *Given a mixture of k Gaussians F such that $\mu_i \neq \mu_j$ for all $i \neq j$, then $3k$ moments suffice to uniquely determine F.*

Other two-parameter distributions also satisfy that $4k-2$ moments are both necessary and sufficient to uniquely determine a k-distribution mixture.

Proposition 7.3. *If F and F' are two mixtures of k uniform distributions[1], then $F = F'$ if and only if $M_j(F) = M_j(F')$ for $1 \leq j \leq 4k-2$.*

[1]We define a uniform mixture as a mixture of k uniform distributions on $[a_i, b_i]$, with $b_i \leq a_{i+1}$ for all i. This is to ensure that it is even possible to distinguish mixtures from their density functions.

The proof of this result is deferred to the appendix; it is very similar to that of Proposition 2.1 in establishing an upper bound by finding a polynomial matching the sign of $f(x) = F - F'$ of degree at most $4k-2$, and establishing a lower bound by finding particular settings for the parameters that make all odd moments 0, as in the Gaussian case.

We suspect the above upper bounds for Gaussians and uniform distributions are similarly true for Laplace and Gamma distributions; we found numerical evidence with Mathematica that 6 moments suffice to determine a mixture of two components for each of these distributions, for example. We conjecture that this is sufficient in the general case of two parameter distributions whose moments are degree k in each of the parameters.

Conjecture 7.4. *Let C be a two-parameter distribution whose moments are degree at most k in each of the parameters. If X and X' are two mixtures of k C-distributions, then $X = X'$ if $M_j(X) = M_j(X')$ for $1 \leq j \leq 4k-2$.*

8. General Mixtures

All of the results up to this point have shown that the number of moments required to learn one or two parameter distributions, given some polynomial condition on their moments in terms of the parameters, is some (linear) function of k independent of the parameters. The following result extends this to a general family of distributions with an arbitrary number of parameters and gives a bound on the number of sufficient moments exponential in k.

Definition 8.1. Let $\mathcal{D}$ denote the class of distribution families of the form $h(\theta)p(x)e^{q(x)\eta(\theta)}$, where $p(x)$ and $q(x)$ are polynomials whose degrees are independent of θ.

Proposition 8.2. *Consider some $F \in \mathcal{D}$ with $deg(p) = d_1 - 1$, $deg(q) = d_2 > 0$. Then $O(d_1 d_2^{2k})$ moments suffice to uniquely determine a mixture of k F-distributions.*

Proof. Let G be a mixture of k F-distributions with parameters $\theta_1, \ldots, \theta_k$, and G' another mixture of k F-distributions with parameters $\theta'_1, \ldots, \theta'_k$. Let $2k = j$, and define $g(x) = G - G' = \sum_{i=1}^{j} p_i(x)e^{q_i(x)}$, where we incorporate the parameters as constants in the polynomials, since they are fixed for each component. We first prove that $g(x)$ has at most $O(d_1 d_2^j)$ by induction on j; it is clear in the base case $j = 0$. Now, assume it is true for $j-1$, and rewrite $g(x) = p_{2k}(x)e^{q_{2k}(x)} + \sum_{i=1}^{2k-1} p_i(x)e^{q_i(x)}$. This is 0 if and only if $g_2(x) = p_{2k}(x) + \sum_{i=1}^{2k-1} p_i(x)e^{q_i(x)-q_k(x)}$ is 0, by quotienting out $e^{q_{2k}(x)}$. Taking d_1 derivatives of $g_2(x)$ and using the fact that the number of zeroes of a function is at most one more than the number of zeroes of its derivative, we get that $g(x)$ has at most d_1 more zeroes than $g_3(x) = \sum_{i=1}^{2k-1} p'_i(x)e^{q_i(x)-q_k(x)}$, where $p'_i(x) = (\frac{d}{dx}(q_i(x) - q_k(x)))^{d_1} \cdot p_i(x)$, which has degree at most $d_1(d_2 - 1) + d_1 = d_1 d_2$. By our inductive step, $g_3(x)$ has at most $O(d_1 d_2 (d_2^{2k-1})) = O(d_1 d_2^{2k})$ zeroes, which means that $g(x)$ has at most $d_1 + O(d_1 d_2^{2k}) = O(d_1 d_2^{2k})$ zeroes.

Letting the number of zeroes of $g(x)$ be a, we can construct a polynomial $h(x)$ of degree a that matches the sign of $g(x)$, which implies, following our usual argument, that $\int_x |h(x)g(x)|\mathrm{d}x > 0$ so one of the first $a = O(d_1 d_2^{2k})$ moments of G and G' differ. □

This argument only works when d_1, d_2 are constants independent of the parameters, in which case it gives us an effective upper bound on a sufficient number of moments based solely on the number of components of the mixture. Examples of well-known families that satisfy this property include Gaussians, Maxwell–Boltzmann distributions, and Rayleigh distributions. However, our bound is still not linear or even polynomial in k, unlike all bounds in previous sections; this is not especially surprising, since the techniques used in Proposition 8.2 do not use much about the structure of the distribution, and we suspect that this bound can be improved in many cases.

We conjecture that an effective bound in terms of k should always exist, independent of the parameters, in cases in which the distribution family is an exponential family. We suspect that, for natural exponential families at least, this bound is linear in k.

Conjecture 8.3. *Let F be an exponential family. Then there exists a function $f(k)$ independent of the parameters such that at most $f(k)$ moments suffice to uniquely determine a mixture of k F-distributions.*

9. Future Work

Looking to the future, in the one-parameter domain, completely classifying the zero sets of Schur polynomials is still open, which would help in determining which sets of $2k-1$ moments uniquely determine a mixture. Another question of interest is the number or properties of additional moments necessary in cases in which a set of $2k-1$ moments do not uniquely determine a finite distribution. The hope is to be able to eventually generalize this sufficiency condition involving Schur polynomials to linear combinations of moments, as this would then allow for the implementation of a moment-matching algorithm for mixtures of one-parameter distributions not in $\mathcal{C}$, such as natural exponential families with higher-degree polynomial variance functions.

Furthermore, there is much work to be done with two-parameter families and exponential families in general. It is still open, in the generic case with means separated, whether only $3k$ moments suffice to uniquely determine a Gaussian, or whether $4k-2$ moments are still needed; we conjectured that the former is true. Another possible area of work is generalizing our upper and lower bound techniques to get an effective bound polynomial or linear in k for general two-parameter mixtures, and in fact any effective bound in terms of k in this case and in the general case of exponential families is still open.

Finally, outside of a single specific result for Gaussians in [11], almost no work has been done on establishing sample complexity lower bounds for learning mixtures of a certain distribution family. We expect that such arguments would follow [11] and use distance metrics to turn a condition on the number of moments necessary to uniquely determine a

distribution into a statistical lower bound, but for which families such an argument holds is still a largely unexplored question.

10. Acknowledgments

We would like to thank Amelia Perry for being a wonderful mentor, as well as providing extensive suggestions for and review of this paper. We would also like to thank Professor Ankur Moitra and Professor David Jerison for helpful discussions about the project, as well as Slava Gerovitch and SPUR as a whole for giving us such a rewarding research experience!

References

[1] P. Boettcher, P. Moroni, G. Pisoni, and D. Gianola, "Application of a finite mixture model to somatic cell scores of Italian goats," *Journal of dairy science*, vol. 88, no. 6, pp. 2209–2216, 2005.

[2] I. D. Dinov, "Expectation maximization and mixture modeling tutorial," *Statistics Online Computational Resource*, 2008.

[3] R. Frühwirth and M. Liendl, "Mixture models of multiple scattering: computation and simulation," *Computer physics communications*, vol. 141, no. 2, pp. 230–246, 2001.

[4] K. Tanaka and K. Tsuda, "A quantum-statistical-mechanical extension of Gaussian mixture model," in *Journal of Physics: Conference Series*, vol. 95, p. 012023, IOP Publishing, 2008.

[5] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London. A*, vol. 185, pp. 71–110, 1894.

[6] S. Dasgupta, "Learning mixtures of Gaussians," in *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 634–644, IEEE, 1999.

[7] S. Arora, R. Kannan, *et al.*, "Learning mixtures of separated nonspherical Gaussians," *The Annals of Applied Probability*, vol. 15, no. 1A, pp. 69–92, 2005.

[8] S. Dasgupta and L. J. Schulman, "A two-round variant of EM for Gaussian mixtures," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*, pp. 152–159, Morgan Kaufmann Publishers Inc., 2000.

[9] S. Vempala and G. Wang, "A spectral algorithm for learning mixture models," *Journal of Computer and System Sciences*, vol. 68, no. 4, pp. 841–860, 2004.

[10] A. T. Kalai, A. Moitra, and G. Valiant, "Efficiently learning mixtures of two Gaussians," in *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 553–562, ACM, 2010.

[11] M. Hardt and E. Price, "Sharp bounds for learning a mixture of two gaussians," *arXiv:1404.4997*, 2014.

[12] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium.*, pp. 93–102, IEEE, 2010.

[13] M. Belkin and K. Sinha, "Polynomial learning of distribution families," in *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 103–112, IEEE, 2010.

[14] C. N. Morris, "Natural exponential families with quadratic variance functions," *The Annals of Statistics*, pp. 65–80, 1982.

[15] C. F. Gauss, "Methodus nova integralium valores per approximationem inveniendi," *Commentationes Societatis regiae scientarium Gottingensis recentiores*, vol. 3, pp. 39–76, 1814.

[16] D. Hunter, "The positive-definiteness of the complete symmetric functions of even order," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 82, pp. 255–258, Cambridge Univ Press, 1977.

A. Proofs of Results

We prove here those results from the body of the paper whose proofs were deferred to the appendix for the sake of space.

Proof of Proposition 2.3. Let μ be the uniform distribution on $[0, 1]$, and let $p_0, p_1, p_2, \ldots$ be the sequence of orthogonal polynomials to this distribution. Let $f = p_{2k-1}$, and notice that f is orthogonal to any polynomial of degree $\leq 2k-2$, since such a polynomial can always be written as a linear combination in $p_0, p_1, \ldots, p_{2k-2}$. Now, consider the distribution μ' on $[0, 1]$ with density $g(x) = 1 + \epsilon f(x)$, where ϵ is sufficiently small that $g(x) > 0$ on $[0, 1]$. Notice that this is indeed a valid distribution because

$$\int_0^1 \mathrm{d}\mu' = \int_0^1 \mathrm{d}\mu + \epsilon \int_0^1 f(x)x^0 \mathrm{d}\mu = 1,$$

by orthogonality of f. Similarly, by orthogonality, we get that, for $j \leq 2k-2$, the jth moment of μ' is

$$\mathbb{E}_{y\sim\mu'}[y^j] = \int_0^1 y^j \mathrm{d}\mu' = \int_0^1 y^j \mathrm{d}\mu + \epsilon \int_{-1}^1 y^j f(x) \mathrm{d}\mu = \int_0^1 y^j \mathrm{d}\mu = \mathbb{E}_{y\sim\mu}[y^j],$$

as desired. When $j = 2k-1$, $f(x)$ is no longer orthogonal to y^j, which gives that $\epsilon \int_0^1 y^j f(x) \mathrm{d}\mu(x) \neq 0$ so $\mathbb{E}_{y\sim\mu'}[y^j] \neq \mathbb{E}_{y\sim\mu}[y^j]$. Thus, μ and μ' are two probability distributions which match on their first $2k-2$ moments but have different $(2k-1)$st moments.

From here, we apply Proposition 2.2 to generate $F = I(\mu)$, and $F' = I(\mu')$, the unique finite distributions that match μ and μ' on their first $2k-1$ moments. Since this implies that F and F are finite distributions that match on their first $2k-2$ moments but have different $(2k-1)$st moments, we are done. □

Proof of Theorem 4.4. Let F have points $x_1, x_2, \ldots, x_k$ with weights $p_1, p_2, \ldots, p_k$, F' have points $y_1, y_2, \ldots, y_k$ with weights $q_1, q_2, \ldots, q_k$, and let $f(x) = F - F'$ be the difference in probability densities. We will split our argument into two cases: separation in the points of the mixture and separation in the weights.

Case 1: Separation in Points

We suppose in this case that there is separation in the points of the mixture; in other words, that there exists some point $a \in \{x_1, \ldots, x_k, y_1, \ldots, y_k\}$ such that $|a - x_i| \gtrsim \epsilon_2$ for all $x_i \neq a$, $|a - y_i| \gtrsim \epsilon_2$ for all $y_i \neq a$. Without loss of generality let this point be y_k, and consider, as in Proposition 2.1, the polynomial

$$P_1(x) = \left(\prod_{i=1}^{k}(x - x_i)\right)\left(\prod_{i=1}^{k-1}(x - y_i)\right) = \sum_{j=0}^{2k-1} a_j x^j.$$

We know that this is zero everywhere except for y_k, and also that $q_k \gtrsim \epsilon_1$. Furthermore, by ϵ_1-separation, there can be at most one point x_i such that $y_k - x_i = \Theta(\epsilon_2)$; for all other

points $z \in \{x_1, \ldots, x_k, y_1, \ldots, y_k\}/\{x_i, y_k\}$, $|y_k - z| \gtrsim \epsilon_1$. This gives us that

$$\left| \int_x f(x) P_1(x) \mathrm{d}x \right| = |q_k P_1(y_k)| \gtrsim \epsilon_1^{2k-1} \epsilon_2,$$

as desired.

Case 2: Separation in Weights

In the second case, we do not have separation in the points, so each point x_i must have a unique point y_j such that $|x_i - y_j| = o(\epsilon_2)$; there cannot exist more than one such point for each x_i by the ϵ_1 separation between the points within mixtures. Without loss of generality let $|x_i - y_i| = o(\epsilon_2)$ for all i. Now, we know, because F, F' are (ϵ_1, ϵ_2) standard, that there must be separation in the weights; in other words, there must exist some l such that $|p_l - q_l| = \Theta(\epsilon_2)$; without loss of generality let $l = k$, and let $a = |x_k - y_k|$.

If we picked a polynomial having one of these points as a root, the value of the polynomial at the other would be too small to give our desired moment estimation. With this in mind, consider the point $b = x_i + c \cdot (\mathrm{sgn}(y_i - x_i)\epsilon_2)$ where c is chosen such that $|b - y_i|, |b - x_i| \gtrsim \epsilon_1$ for all $i \neq k$, and $|b - y_k|, |b - x_k| \gtrsim \epsilon_2$, possible by the ϵ_1-separation of F, F' and the $o(\epsilon_2)$ closeness of y_k and x_k. With this established, define the polynomial

$$P_2(x) = \left(\prod_{i=1}^{k-1} (x - x_i) \right) \left(\prod_{i=1}^{k-1} (x - y_i) \right) (x - b) = \sum_{j=0}^{2k-1} b_j x^j.$$

Notice that $P_2(x)$ is zero everywhere except for x_k and y_k. We want to evaluate $p_k P(x_k) - q_k P(y_k)$, since the absolute value of this is exactly the integral $|\int_x P_2(x) f(x) dx|$. Without loss of generality let $p_k > q_k$; in particular, we know that $p_k - q_k \gtrsim \epsilon_2$. From this, notice that $|\frac{p_k}{q_k}| \geq 1 + \epsilon_2$. We now want to show that $\frac{P(y_k)}{P(x_k)} \ll 1 + \epsilon_2$, which will also us to bound $|\int_x P_2(x) f(x) dx|$. Notice that each term of $P_2(x_k)$ and $P_2(y_k)$, excepting $(x - b)$, are $\Omega(\epsilon_1)$, and in particular differ by a. Thus, the multiplicative factor by which these terms differ is at most $\frac{\epsilon_1 + a}{\epsilon_1}$, so in total, excluding the $x - b$ term, the multiplicative factor is at most $(1 + \frac{a}{\epsilon_1})^k \asymp 1 + ka$, since $\epsilon_1 \gg a$. Finally, notice that because b is chosen to be farther from x_i, $(x_i - b)$ is larger than $(y_i - b)$, the maximum multiplicative factor that this contributes to $\frac{P(y_k)}{P(x_k)}$ is 1. Thus, we have that $P(y_k) \gtrsim (1 + ka) P(x_k)$, so

$$p_k P(x_k) - q_k P(y_k) \gtrsim q_k P(x_k)((1 + \epsilon_2) - (1 + ka)) \gtrsim \epsilon_2 q_k P(x_k).$$

Since $P(x_k) \gtrsim \epsilon_1^{2k-2} \epsilon_2$ and $q_k \gtrsim \epsilon_1$, this gives that

$$\left| \int_x f(x) P_2(x) \mathrm{d}x \right| \gtrsim \epsilon_2 \epsilon_1^{2k-1},$$

as in the first case.

Using Case 1 and Case 2 to find differing moments

From our two cases, we have that, regardless of the closeness of the points, we have that there exists a polynomial $P_i(x)$ such that $|\int_x P_2(x) f(x) dx| \gtrsim \epsilon_1^{2k-1} \epsilon_2$. Now, notice that each

coefficient a_j and b_j of $P_i(x)$ are $O(\epsilon_1^{-2k+1})$, since each polynomial is of degree $2k-1$ and the x_i, y_i are bounded above by $\frac{1}{\epsilon_1}$. Thus, we have that

$$\left|\int_x P_2(x)f(x)dx\right| \lesssim \epsilon_1^{-2k+1}\left|\int_x x^i f(x)dx\right|,$$

which gives, by our previous bounds, that

$$\left|\int_x x^i f(x)dx\right| = \left|\sum_{i=1}^{k} M_i(F) - M_i(F')\right| \gtrsim \epsilon_1^{4k-2}\epsilon_2.$$

This implies that there must exist some i such that $|M_i(F) - M_i(F')| > \Omega(\epsilon_1^{4k-2}\epsilon_2)$, as desired. □

Proof of Proposition 4.6. The proof of this is almost identical to that of Lemma 3.2 in [11]. Let $s = \frac{2}{\epsilon^2(2t)^{2t-1}} = O(\frac{1}{\epsilon^2})$, and partition the samples x_i into groups of size s. Consider taking the empirical moment of each group. Because F is sub-exponential, $M_p(F) \leq p^p\lambda^p$ by definition, so we have that

$$\mathrm{Var}(x_i^p) \leq \mathbb{E}(x_i^{2p}) \leq (2p)^{2p}\lambda^{2p}$$

Now, consider the empirical pth moment of a group $\widehat{M_p} = \frac{1}{s}\sum_{i=1}^{s} x_i^p$; the above bound gives that $\mathrm{Var}\left(\widehat{M_p}\right) \leq \frac{(2p)^{2p}\lambda^{2p}}{s}$.

By Chebyshev's inequality, we then have that

$$\mathbb{P}\left(\left|\widehat{M_p} - M_p\right| > \frac{\lambda^p(2p)^p}{\sqrt{cs}}\right) \leq c.$$

Letting $c = \frac{1}{4t}$ and plugging in for s gives

$$\mathbb{P}\left(\left|\widehat{M_p} - M_p\right| > \epsilon\lambda^p\right) \leq \frac{1}{4t}$$

By a union bound in each group this gives

$$\mathbb{P}\left(\left|\widehat{M_p} - M_p\right| \leq \epsilon\lambda^p\right) \geq \frac{3}{4}$$

for all $p \leq t$. From this, because there are $O(\log(1/\delta))$ groups, the probability that more than half of the groups satisfy $\mathbb{P}\left(\left|\widehat{M_p} - M_p\right| \leq \epsilon\lambda^p\right)$ is at most $1-\delta$, which means that, if we take the median of our estimated moments for each of the groups, it will satisfy the desired bound, and we are done. □

Proof of Proposition 7.3. We first prove that $4k-2$ moments is an upper bound, the if direction of this proof. To do this, assume $F \neq F'$, and let $f(x) = F - F'$ be the difference in probability densities. We will show that there exists a polynomial $p(x)$ of degree at most $4k-2$ such that $\int_x |p(x)f(x)| > 0$. To do this, consider the zeroes of $f(x)$; these can only occur when the value of $f(x)$ changes, which occurs only at the left or right edge of one of the uniform distributions in one of the mixtures. There are $2k$ such distributions, so there are $4k$ changes in value in total. However, two of these, the right and leftmost, do not entail a sign change, since past each of these $f(x) = 0$ for all x. Let these two boundary points be a and b, respectively, and, With this in mind, let the set of zeroes of $f(x)$ between a and b be $r_1, r_2, \ldots, r_j$, where $j \leq 4k-2$. Then, defining $p(x) = c(x-r_1)(x-r_2)\cdots(x-r_j) = \sum_{i=0}^{4k-2} \alpha_i x^i$, we can pick $c \in \{-1, 1\}$ such that $p(x)$ matches the sign of $f(x)$ at all x. Since we cannot have $f(x) = 0$ for all x because $F \neq F'$, we then have that $\int_x |p(x)f(x)| > 0$, as desired. This is equivalent to saying that $\sum_{i=1}^{4k-2} \alpha_i \int_x |x^i f(x)| > 0$, which implies that $\int_x |x^j f(x)| \neq 0$ for some j, or that $M_j(F) \neq M_j(F')$. Thus, if $F \neq F'$, some pair of the first $4k-2$ moments differs, so $4k-2$ moments are indeed sufficient to determine a mixture of k uniform distributions.

The only if direction proceeds similarly as in the Gaussian case. Let the uniform distributions of F have parameters $(a_1, b_1), (a_2, b_2), \ldots, (a_k, b_k)$ and those of F' have parameters $(c_1, d_1), (c_2, d_2), \ldots, (c_k, d_k)$. Now, let $a_i = -b_i$ and $c_i = -d_i$. Using that the moments of a uniform distribution with parameters (a, b) are $M_n = \frac{b^{n+1} - a^{n+1}}{(n+1)(b-a)}$, this makes all odd moments of the uniform distributions in F, F' to be 0, which means that all odd moments of F, F' are zero, since they are just convex combinations of their components. The even moments, in turn, are b_i^{2n} and d_i^{2n}, so the even moments of F are $M_{2n}(F) = \sum_{i=1}^{k} p_i b_i^{2n}$, and similarly for F'. This gives that $M_{2n}(F)$ is equal to the nth moment of a finite distribution with points at b_i^2 and weights p_i; by Proposition 2.3 we can match the first $2k-2$ moments of this distribution by picking corresponding positive d_i^2 (as in the Gaussian case, we can ensure positivity by taking our interval to be $[0, 1]$, which by Proposition 2.2 implies that our d_i^2 fall in that range), implying that the even moments of F, F' match up to $4k-4$; since we know from our choice of a_i that the odd moments match since they are all 0, we have a pair of F, F' such that the first $4k-3$ moments match, implying our lower bound. $\square$

B. Natural Exponential Families with Quadratic Variance Functions

In Table 1 below, we give the probability density function, variance function, cumulants, and moment recurrence for (up to affine transformations, division, and convolution) five of the NEF-QVF's, omitting the generalized hyperbolic secant distribution due to its complicated and esoteric nature. Excluding the binomial distribution, these are examples of well-known distribution whose mixtures that can be learned efficiently by Algorithm 4.1.

Distribution	PDF	Variance Function	Cumulants or Cumulant Recurrence	Moments or Moment Recurrence
Gaussian with fixed variance	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$V(\mu) = \sigma^2$	$C_1 = \mu$, $C_2 = \sigma^2$, $C_k = 0 \quad \forall k \geq 3$	$M_k = \mu M_{k-1} + (k-1)\sigma^2 M_{k-2}$
Poisson distribution	$P(x) = \frac{\lambda^k}{k!} e^{-\lambda}$	$V(\mu) = \mu$	$C_k = \mu \quad \forall k \geq 1$	$M_k = \mu M_{k-1} + \mu \frac{d(M_{k-1})}{d\mu}$
Gamma Distribution with Fixed Shape	$f(x) = \frac{x^{r-1}}{\beta^r \Gamma(r)} e^{-\frac{x}{\beta}}$	$V(\mu) = \frac{\mu^2}{r}$	$C_k = \frac{(k-1)!}{r} \mu^k$	$M_k = \prod_{j=0}^{k-1} (r+j) \left(\frac{\mu}{r}\right)^k$
Binomial Distribution with Fixed # of Trials	$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$	$V(\mu) = -\frac{1}{n}\mu^2 + \mu$	$C_k = \left(-\frac{\mu^2}{n} + \mu\right) C_k'(\mu)$	$M_k = \mu M_{k-1} + \frac{\mu(n-\mu)}{n} M_{k-1}'(\mu)$
Negative Binomial Distribution with Fixed Failure Number	$P(x) = \frac{\Gamma(x+r)}{\Gamma(r)x!} p^x (1-p)^r$	$V(\mu) = \frac{1}{r}\mu^2 + \mu$	$C_k(\mu) = \left(\frac{1}{r}\mu^2 + \mu\right) C_{k-1}'(\mu)$	$M_k(\mu) = \mu M_{k-1} + \mu \left(\frac{\mu}{r} + 1\right) M_k'(\mu)$

Table 1: The probability distribution functions, variance functions, cumulants, and moments of five of the six NEF-QVF's.