Polynomial Methods in Statistical Inference: Theory and Practice

Other titles in Foundations and Trends® in Communications and Information Theory

Cache Optimization Models and Algorithms
Georgios Paschos, George Iosifidis and Giuseppe Caire
ISBN: 978-1-68083-702-5

Lattice-Reduction-Aided and Integer-Forcing Equalization: Structures, Criteria, Factorization, and Coding
Robert F. H. Fischer, Sebastian Stern and Johannes B. Huber
ISBN: 978-1-68083-644-8

Group Testing: An Information Theory Perspective
Matthew Aldridge, Oliver Johnson and Jonathan Scarlett
ISBN: 978-1-68083-596-0

Sparse Regression Codes
Ramji Venkataramanan, Sekhar Tatikonda and Andrew Barron
ISBN: 978-1-68083-580-9

Polynomial Methods in Statistical Inference: Theory and Practice

Yihong Wu
Department of Statistics and Data Science
Yale University
USA
yihong.wu@yale.edu

Pengkun Yang
Department of Electrical Engineering
Princeton University
USA
pengkuny@princeton.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

Y. Wu and P. Yang. *Polynomial Methods in Statistical Inference: Theory and Practice*. Foundations and Trends® in Communications and Information Theory, vol. 17, no. 4, pp. 402–585, 2020.

ISBN: 978-1-68083-731-5

Foundations and Trends® in Communications and Information Theory

Volume 17, Issue 4, 2020

Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2020, Volume 17, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328 . Also available as a combined paper and online subscription.

Contents

Polynomial Methods in Statistical Inference: Theory and Practice

Yihong Wu[1] and Pengkun Yang[2]

[1] *Department of Statistics and Data Science, Yale University, USA; yihong.wu@yale.edu*

[2] *Department of Electrical Engineering, Princeton University, USA; pengkuny@princeton.edu*

ABSTRACT

This survey provides an exposition of a suite of techniques based on the theory of polynomials, collectively referred to as polynomial methods, which have recently been applied to address several challenging problems in statistical inference successfully. Topics including polynomial approximation, polynomial interpolation and majorization, moment space and positive polynomials, orthogonal polynomials and Gaussian quadrature are discussed, with their major probabilistic and statistical applications in property estimation on large domains and learning mixture models. These techniques provide useful tools not only for the design of highly practical algorithms with provable optimality, but also for establishing the fundamental limits of the inference problems through the method of moment matching. The effectiveness of the polynomial method is demonstrated in concrete problems such as entropy and support size estimation, distinct elements problem, and learning Gaussian mixture models.

Yihong Wu and Pengkun Yang (2020), "Polynomial Methods in Statistical Inference: Theory and Practice", Foundations and Trends® in Communications and Information Theory: Vol. 17, No. 4, pp 402–585. DOI: 10.1561/0100000095.

1

Introduction

Modern data-analytic applications frequently involve complex and high-dimensional statistical models. For example, applications such as natural language processing, genetics, and neuroscience deal with datasets naturally viewed as being sampled from *probability distributions over a large domain*. A number of real-world signal processing and machine learning tasks rest upon data-driven procedures for estimating *distributional properties* (functionals of the data-generating distribution), including *entropy* for understanding the neural coding [7, 9, 65, 112, 141, 168, 183], *mutual information* for image registration in fMRI [118, 157, 187, 188] and learning graphical models [40, 99], etc. For these tasks, the key challenge is to accurately estimate the property even when the domain size far exceeds the sample size and the distribution itself is impossible to learn.

Another prominent example of complex statistical models deals with *mixture models*, which are useful to model the effects of latent variables and form the basis of many clustering algorithms. The simplest mixture models is perhaps the Gaussian mixture model, introduced by Pearson in 1894 to model the presence of hidden subpopulations within an overall population. Despite the seemingly innocuous nature of the Gaussian mixture models, many difficult challenges arise, such as the vanishing Fisher information leading to nonparametric rates, the nonexistence of maximum likelihood estimator in location-scale mixtures, etc. For this reason, it proves to be a fertile ground for innovations in statistical methodologies, including the method of moments [151], the Expectation-Maximization (EM) algorithm [49], the Generalized Method of Moments [82], etc. Despite the vast literature and recent breakthroughs, many problems as basic as optimal estimation rates remain open in finite mixture models.

Recently, several challenging problems in property estimation and mixture models have been successfully resolved using methods based on the theory of polynomials, in particular, polynomial approximation, interpolation, as well as moments and positive polynomials. They provide useful tools not only for the design of algorithms that are both statistically optimal and computationally efficient, but also in establishing the fundamental limits of the inference problems. This survey aims to provide an exposition of these techniques, which are collectively referred to as the *polynomial method*, as well as their application in statistical inference.

1.1 Background on Polynomial Methods

The theory of polynomials is a rich subject in mathematics of both algebraic and analytic flavor. It forms the foundation of and has diverse applications in many subjects including optimization, combinatorics, coding theory, control theory, digital signal processing, game theory, statistics and machine learning, etc, leading to many deep theoretical results and highly practical algorithms. In this survey, we mainly focus on polynomial approximation, interpolation, and positive polynomials that will be introduced below.

Polynomial Approximation and Interpolation. One of the most well-understood subjects in approximation theory, polynomial approximation aims at approximating a given complicated function, in either a local or global sense, using algebraic or trigonometric polynomials of a certain degree. For instance, the Taylor expansion characterizes the local behavior of a smooth function and provide the foundation for optimization techniques such as gradient descent and the Newton-Raphson method [136] and kernel-based methods in statistical inference [83, 190]; trigonometric polynomials represent functions in the frequency domain through Fourier analysis, which are the theoretical underpinnings for digital signal processing and wireless transmission [144, 189]. A closely related topic is polynomial interpolation, which can be viewed as achieving zero approximation error on a discrete set of points.

In property estimation, the functional to be estimated can be highly nonsmooth and classical methods requires a large sample size in order to be accurate. In such settings, polynomial approximation and interpolation provide a useful primitive for constructing better estimates by first approximating the original functional by a polynomial and then estimate the polynomial approximant. Besides the approximation error which is the primary concern in approximation theory, other properties of the polynomial approximant such as the magnitude of its coefficients are also crucial for bounding the statistical error.

Moments and Positive Polynomials. The theory of moments plays a key role in the developments of analysis, probability, statistics, and optimization. We refer the readers to the classics [106, 117, 177] and the more recent monographs [120, 174] for a detailed treatment. In statistical inference, the method of moments was originally introduced by Pearson [151] for mixture models, which constructs estimates by solving polynomial equations. Due to its conceptual simplicity and flexibility, especially in models without the complete specification of the joint distribution of data, method of moments and its extensions have been widely applied in practice, for instance, to analyze economic and financial data [76]. In probability and optimization literature, the classical moment problem refers to determining whether a probability distribution is determined by all of its moments. Solution to the moment problem requires understanding the moment space, which is the convex set formed by moments of probability distributions. The moment space satisfies many geometric properties (such as the Cauchy-Schwarz and Hölder inequalities) and a complete description can be phrased in terms of positive polynomials, which are further related to sums of squares and semidefinite programming. Together with techniques based on polynomial interpolation, this structural information can be leveraged to design moment-based methods for learning mixture models that are statistically optimal, robust to model misspecification, and highly practical.

1.2 Polynomial Methods for Designing Estimators

We will apply the above polynomial methods to the tasks of estimating distributional properties and learning mixture models with the goal of constructing estimators with good statistical performance.

Estimating Distributional Properties on Large Domains. Given samples drawn from an unknown distribution P on a large domain, the goal is to estimate a specific property of that distribution, such as various information measures including the Shannon entropy, Rényi entropy, and the support size. This falls under the category of *functional estimation* [164], where we are not interested in directly estimating the high-dimensional parameter (the data-generating distribution P) per se, but rather a function thereof. Estimating a distributional functional has been intensively studied in nonparametric statistics, including estimating a scalar function of a regression function or density such as linear functionals [55, 181], quadratic functionals [33, 121], L_q norm [123], etc.

To estimate a functional, perhaps the most natural idea is the "plug-in" approach, namely, first estimate the parameter and then substitute into the function. As frequently observed in the functional estimation literature, the plug-in estimator can suffer from severe bias (see [21, 60] and the references therein). Indeed, although the plug-in estimate is typically asymptotically efficient and minimax (cf., e.g., [199, Sections 8.7 and 8.9]) for fixed domain size, it can be highly suboptimal in high dimensions, where, due to the large alphabet and resource constraints, we are constantly contending with the difficulty of *undersampling* in applications such as

- Natural language processing: The vast vocabulary size of natural languages, compounded by the frequent use of bigrams and trigrams in practice [131], leads to an effective alphabet size far exceeding the sample size. A well-known example from corpus linguistics is that about half of the words in the Shakespearean canon only appeared once [59];

- Neuroscience: in analyzing neural spike trains, natural stimuli generate neural responses of high timing precision resulting in a massive space of meaningful responses [22, 130, 172];

- Network traffic analysis: many customers or website users are only seen a small number of times [20].

Statistical inference on large domains has a rich history in information theory, statistics and computer science, with early contributions dating back to Fisher, Good and Turing, Efron and Thisted, etc. [59, 62, 73, 185] and recent renewed interests on compression, prediction, classification and estimation on large alphabets [23, 109, 145, 198, 204]; however, none of the aforementioned results allows a general understanding of the fundamental limits of estimating information quantities of large distributions. While there exists a vast literature on information-theoretic approaches to the statistical inference of high-dimensional parameters [24, 93, 122, 155, 216, 217], a systematic theory for estimating their low-dimensional functionals remains severely under-developed, especially in the *sublinear regime* where the sample size is far less than the domain size so that the underlying distribution is impossible to learn but certain low-dimensional features can nevertheless be estimated accurately.

In this survey, we will investigate a few prototypical problems in estimating distributional properties such as the Shannon entropy and the support size. These properties can be easily estimated if the sample size far exceeds the support size of the underlying distribution, but how can it be done if the observations are relatively scarce, especially in the *sublinear regime* where the sample size is far less than the domain size? It turns out the theory of polynomial approximation provides a principled approach to construct an optimal estimator. To illustrate this program let us consider the problem of estimating a function $f(p)$ based on n independent observations drawn from Bernoulli distribution with mean p, or equivalently, the sufficient statistic $N \sim \text{Binomial}(n, p)$. This simple setting forms the basis of designing estimators for distributional properties in Sections 3–5. Given any estimator $\hat{f}(N)$, its mean is given by

$$\mathbb{E}[\hat{f}(N)] = \sum_{j=0}^{n} f(j) \binom{n}{j} p^j (1-p)^{n-j},$$

which is a degree-n polynomial in p. Consequently, unless the function f is a polynomial, there exists no unbiased estimator for $f(p)$. Conversely, given any degree-n polynomial $\tilde{f}$, we can always construct an unbiased estimator for $\tilde{f}(p)$ by combining the unbiased estimator of each monomial (see, e.g., (3.9) in Subsection 3.2). These observations suggest that, for the purpose of reducing the bias, we should first find a polynomial $\tilde{f}$ of degree at most n such that the approximation error $|f(p) - \tilde{f}(p)|$ is small for every possible values of p, and then construct an unbiased estimator $\hat{f}(N)$ for $\tilde{f}(p)$. Fixing $L \leq n$, the best degree-L polynomial $\tilde{f}$ that minimizes the worst-case approximation error can be found by solving the following optimization problem:

$$\inf_{\lambda_0, \ldots, \lambda_L} \sup_{p} \left| f(p) - \sum_{i=0}^{n} \lambda_i p^i \right|; \tag{1.1}$$

this is known as the *best uniform polynomial approximation* problem which will be discussed at length in Subsection 2.1. Although the approximation error decays with the degree, typically we cannot choose it to be as large as n since the estimation error of monomials grows rapidly with the degree. Therefore, the degree L must be chosen appropriately (often logarithmic in the sample size n) so as to balance the approximation error and the estimation error (the bias-variance tradeoff). This method was pioneered by Lepski *et al.* [123] for nonparametric regression and further developed in Cai and Low [34] for the Gaussian sequence model. We will elaborate on the high-level ideas in Section 3 and illustrate the effectiveness of this approach in Sections 4 and 5 for specific problems.

Learning Gaussian Mixtures. Sampling from a mixture model can be viewed as being a two-step process: first draw a latent parameter $\theta \sim \nu$; then draw an observation $X \sim P_\theta$. The marginal distribution of each sample is

$$\pi_\nu = \int P_\theta \mathrm{d}\nu(\theta). \tag{1.2}$$

We refer to ν as the mixing distribution and π_ν as the mixture distribution. A finite mixture model has a discrete mixing distribution of finite support and a mixture distribution of the form $\sum_i w_i P_{\theta_i}$. The key question in mixture model is the following: If we are only given unlabeled data from the mixture model, can we reconstruct the parameters in each component accurately and efficiently? Furthermore, in the regime where it is impossible to learn the labels with small misclassification rate, is it still possible to learn the mixing distribution and the mixture distribution accurately?

In the special case that each P_θ is a Gaussian distribution, this is the problem of learning Gaussian mixtures, a classical problem in statistics dating back to the work of Pearson [151]. In addition, methods for learning Gaussian mixtures are widely used as part of the core machine learning toolkit, such as the popular scikit-learn package in Python [152], Google's Tensorflow [1], and Spark's MLlib [133]; however, few provable guarantees are available. It is only recently proved in [104, 138] that a mixture of constant number of components can be learned in polynomial time using a polynomial number of observations. The optimal rate for learning finite Gaussian location mixtures is recently determined in [56, 86, 211] and for location-scale mixture only for the special case of two components [84]. Is there a systematic way to obtain the sharp error rates and how to efficiently and optimally learn a Gaussian mixture? We will investigate the moment methods for the optimal estimation of Gaussian mixtures, where we learn a discrete mixing distribution by learning its moments. The key observation is that as opposed to the vanilla method of solving moment equations, the moment estimates should be first denoised based on the geometry of the moment space, and the denoising step can be efficiently carried out through convex optimization (semidefinite programming). The learned moments can be then converted to a discrete distribution by the efficient algorithm of Gaussian quadrature. This approach will be presented in Sections 6–7.

1.3 Polynomial Methods for Determining Theoretical Limits

Another focus of this survey is to investigate the fundamental limits of statistical inference, that is, the optimal estimation error among all estimators regardless of computational costs. While the use of polynomial methods on the constructive side is admittedly natural, the fact that it also arises in the optimal lower bound is perhaps surprising.

To give a precise definition of the fundamental limits, we begin with an account of the general framework for statistical inference. We assume that the sample $X_1, \ldots, X_n$ are independently generated from an unknown distribution P that belongs to a collection of distributions $\mathcal{P}$. The goal is to estimate a certain property $T(P)$ of the distribution P.

In this survey we consider the following two types of problems:

- *Estimating distributional properties*: $T(P)$ is a functional of the unknown discrete distribution $P = (p_1, p_2, \ldots)$, such as the Shannon entropy

  $$H(P) = \sum_i p_i \log \frac{1}{p_i} \tag{1.3}$$

  and the support size

  $$S(P) = \sum_i \mathbf{1}_{\{p_i > 0\}}. \tag{1.4}$$

- *Learning Gaussian mixtures*: P is a Gaussian mixture and $T(P)$ represents the parameters, including the mean, variance, and the mixing weights, of each Gaussian component. Equivalently, $T(P)$ can be viewed as the mixing distribution of the mixture model (see Section 6).

Given a loss function $\ell(\hat{T}, T(P))$ that measures the accuracy of an estimator $\hat{T}$, the decision-theoretic fundamental limit is defined as the *minimax risk*

$$R_n^* \triangleq \inf_{\hat{T}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\hat{T}, T(P))], \tag{1.5}$$

where the infimum is taken over all estimators $\hat{T}$ measurable with respect to $X_1, \ldots, X_n$ drawn independently from P. Examples of the loss function include the quadratic loss $\ell(x,y) = \|x-y\|_2^2$ and the zero-one loss $\ell(x,y) = \mathbf{1}_{\{\|x-y\|_2>\epsilon\}}$ for a desired accuracy ϵ. For the zero-one loss, we also consider the *sample complexity*:

Definition 1.1. For a desired accuracy ϵ and confidence $1-\delta$, the sample complexity is the minimal sample size n such that there exists an estimator $\hat{T}$ based on n independent and identically distributed (i.i.d.) observations drawn from a distribution P such that $\mathbb{P}[\ell(\hat{T}, T(P)) < \epsilon] \geq 1-\delta$ for any $P \in \mathcal{P}$.

In this survey, our primary goal is to characterize the minimax risk (1.5) within universal constant factors, which is known as the *minimax rate*; we will also consider the sample complexity in Definition 1.1. This task entails an upper bound achieved by certain estimators, preferably a computationally efficient one, and a matching minimax lower bound that applies to all estimators.

A general program for obtaining lower bounds is based on a reduction of estimation to testing (Le Cam's method); cf. Subsection 3.3. If there are two distributions P and Q that cannot be reliably distinguished based on a given number of independent observations, while $T(P)$ and $T(Q)$ are different, then any estimate suffers a maximum risk at least proportional to the distance between $T(P)$ and $T(Q)$. Furthermore, sometimes one needs to consider a pair of randomized distributions in which case one needs to construct two distributions (priors) on the space of distributions (also known as fuzzy hypothesis testing in [190]). Here the polynomial method enters the scene again: statistical closeness between two distributions can be bounded by comparing their moments. More precisely, the strategy is to choose two priors with matching moments up to a certain degree, which ensures the induced distributions of data are impossible to test. The minimax lower bound is then given by the maximal separation in the expected functional values subject to the moment matching condition. For example, it pertains to the optimal value of the following type of moment matching problem:

$$\begin{aligned} \sup \quad & \mathbb{E}_\nu[f(X)] - \mathbb{E}_{\nu'}[f(X)], \\ \text{s.t.} \quad & \mathbb{E}_\nu[X^j] = \mathbb{E}_{\nu'}[X^j], \quad j = 0, \ldots, L, \\ & \nu, \nu' \text{ are supported on } [a,b], \end{aligned} \tag{1.6}$$

where the supremum is over all pairs of distributions, and the function f, the degree L, and the interval $[a, b]$ are problem specific. We will discuss how to choose those parameters, construct a pair of least favorable priors from the optimal solution, and then derive the minimax lower bound in Sections 4 and 5. It turns out this optimization problem is the *dual* problem of the best polynomial approximation that arises in the design of polynomial-based estimator in Subsection 1.2. In the introduction, let us first look into the relation to polynomial method. Below we formally derive the duality, and we leave the discussion on strong duality and the correspondence between primal and dual solutions to Subsection 2.2. By introducing the Lagrangian multipliers $\lambda_1, \ldots, \lambda_L$, we optimize the Lagrangian function by

$$\begin{aligned}
&\sup_{\nu,\nu'} \mathbb{E}_\nu[f(X)] - \mathbb{E}_{\nu'}[f(X)] - \sum_{j=1}^{L} \lambda_i (\mathbb{E}_\nu[X^j] - \mathbb{E}_{\nu'}[X^j]) \\
&= \sup_{\nu,\nu'} \mathbb{E}_\nu \left[f(X) - \sum_{j=1}^{L} \lambda_i X^i \right] - \mathbb{E}_{\nu'} \left[f(X) - \sum_{j=1}^{L} \lambda_i X^i \right] \\
&= \sup_{x \in [a,b]} \left(f(x) - \sum_{j=1}^{L} \lambda_i x^i \right) - \min_{x \in [a,b]} \left(f(x) - \sum_{j=1}^{L} \lambda_i x^i \right).
\end{aligned}$$

We can introduce another variable λ_0 that does not impact the optimal value and formulate the dual problem as

$$\begin{aligned}
&\inf_{\lambda_0,\ldots,\lambda_L} \sup_{x \in [a,b]} \left(f(x) - \sum_{j=0}^{L} \lambda_i x^i \right) - \min_{x \in [a,b]} \left(f(x) - \sum_{j=0}^{L} \lambda_i x^i \right) \\
&= 2 \inf_{\lambda_0,\ldots,\lambda_L} \sup_{x \in [a,b]} \left| f(x) - \sum_{j=0}^{L} \lambda_i x^i \right|. \qquad (1.7)
\end{aligned}$$

This last formulation is precisely the best polynomial approximation problem (1.1). For this reason, estimators constructed using the method of polynomial approximation frequently comes with a matching lower bound that certifies their statistical optimality. The connection is precisely the duality between polynomial approximation and moment matching.

The method of moment matching can be similarly carried out for learning mixture models. Typically, there is a minimal number L of moments that identifies a finite mixture model, which depends on the order (the number of components) of the mixture model. A statistical lower bound can then be obtained by constructing a pair of distributions with matching $L-1$ moments. This naturally matches the performance of the "most economical" moment-based estimators that learns the mixture distribution using the minimal number of moments. We will discuss this approach in Section 7.

1.4 Organization

In this survey, we present several tools from the theory of polynomials and their applications in statistical problems. Section 2 provides a brief introduction to the necessary background in the theory of polynomials, including polynomial approximation, interpolation and majorization, theory of moments and positive polynomials, orthogonal polynomials, and Gaussian quadrature. Figure 1.1 describes how these techniques are used in specific statistical applications.

The first statistical application is in the topic of property estimation. Section 3 introduces some common framework and techniques, including Poisson sampling, approximation-theoretic construction of statistical

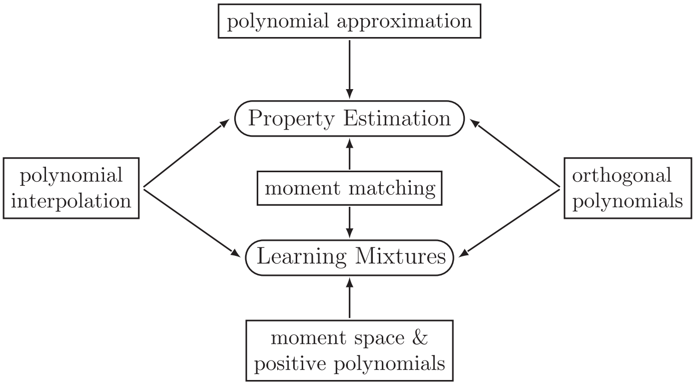

Figure 1.1: Statistical applications of polynomial methods.

estimators, and minimax lower bounds based on moment matching. We then apply these techniques to two representative problems: The problem of entropy estimation is studied in details in Section 4; In Section 5, we study the estimation of the unseen, including estimating the support size and the distinct elements problem.

The second statistical application is learning Gaussian mixture models using moment methods. A general framework for mixture models and various moment comparison theorems are developed Section 6, which form the underpinnings of our statistical theory. Most of these results do not depend on properties of Gaussians and are applicable to general mixture models. Section 7 describes algorithms for Gaussian mixture models and their statistical guarantees, complemented by matching lower bounds.

1.5 Notations

For $k \in \mathbb{N}$, let $[k] \triangleq \{1, \ldots, k\}$. We use standard big-$O$ notations, e.g., for any positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \leq Cb_n$ for some absolute constant $C > 0$, $a_n = o(b_n)$ or $a_n \ll b_n$ or if $\lim a_n/b_n = 0$. We write $o_\delta(1)$ as $\delta \to 0$ to indicate convergence that is uniform in all other parameters. The notations $a \wedge b$ and $a \vee b$ stand for $\min\{a, b\}$ and $\max\{a, b\}$, respectively. For a probability measure π on the real line, let F_π denote its cumulative distribution function (CDF), with $F_\pi(t) \triangleq \pi((-\infty, t])$. A distribution π is called σ-subgaussian if $\mathbb{E}_\pi[e^{tX}] \leq \exp(t^2\sigma^2/2)$ for all $t \in \mathbb{R}$. For matrices $A \succeq B$ stands for $A - B$ being positive semidefinite. The Euclidean ball centered at $x \in \mathbb{R}^d$ of radius r is denoted by $B(x, r)$.

Denote by $\mathrm{Binomial}(n, p)$ the binomial distribution with n Bernoulli trials and success probability p. For $P = (p_1, \ldots, p_k)$, denote by $\mathrm{Multinomial}(n, P)$ the multinomial distribution with n trials where each trial has outcome i with probability p_i. Denote by $N(\mu, \sigma^2)$ the normal distribution with mean μ and variance σ^2 and let $\phi(x) \triangleq \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ denote the standard normal density. Denote by $\mathrm{Poi}(\mu)$ the Poisson distribution with mean μ.

We recall the definition of the following f-divergences (cf. [190, Chap. 2] for details). For probability distributions P and Q, the Kullback-Leibler (KL) divergence is $D(P\|Q) \triangleq \int dP \log \frac{dP}{dQ}$ if $P \ll Q$ and ∞ otherwise; the χ^2-divergence is defined as $\chi^2(P\|Q) \triangleq \int dP(\frac{dP}{dQ} - 1)^2$ if $P \ll Q$ and ∞ otherwise; the squared Hellinger distance is $H^2(P,Q) \triangleq \int(\sqrt{\frac{dP}{d\mu}} - \sqrt{\frac{dQ}{d\mu}})^2 d\mu$ and the total variation distance is $\mathsf{TV}(P,Q) \triangleq \int |\frac{dP}{d\mu} - \frac{dQ}{d\mu}| d\mu$, for any dominating measure μ such that $P \ll \mu$ and $Q \ll \mu$.

References

[1] Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng (2016). "Tensorflow: A system for large-scale machine learning". In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. USENIX Association. 265–283.

[2] Abramowitz, M. and I. A. Stegun (1964). *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Courier Corporation.

[3] Acharya, J., H. Das, A. Orlitsky, and A. T. Suresh (2017). "A unified maximum likelihood approach for estimating symmetric properties of discrete distributions". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR. 11–21.

[4] Acharya, J., A. Orlitsky, A. T. Suresh, and H. Tyagi (2015). "The complexity of estimating Rényi entropy". In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM. 1855–1869.

[5] Achlioptas, D. and F. McSherry (2005). "On spectral learning of mixtures of distributions". In: *Proceedings of the 18th Annual Conference on Learning Theory (COLT 2005)*. Berlin, Heidelberg: Springer. 458–469.

[6] Akhiezer, N. I. (1965). *The Classical Moment Problem: And Some Related Questions in Analysis*. Vol. 5. Oliver & Boyd.

[7] Aktulga, H. M., I. Kontoyiannis, L. A. Lyznik, L. Szpankowski, A. Y. Grama, and W. Szpankowski (2007). "Identifying statistical dependence in genomic sequences via mutual information estimates". *EURASIP Journal on Bioinformatics and Systems Biology*. 2007: 3.

[8] Améndola, C., K. Ranestad, and B. Sturmfels (2016). "Algebraic identifiability of Gaussian mixtures". *International Mathematics Research Notices*. 2018(21): 6556–6580.

[9] Amigó, J. M., J. Szczepański, E. Wajnryb, and M. V. Sanchez-Vives (2004). "Estimating the entropy rate of spike trains via Lempel-Ziv complexity". *Neural Computation*. 16(4): 717–736.

[10] Andrews, D. F. and C. L. Mallows (1974). "Scale mixtures of normal distributions". *Journal of the Royal Statistical Society: Series B (Methodological)*. 36(1): 99–102.

[11] Arora, S. and R. Kannan (2001). "Learning mixtures of arbitrary Gaussians". In: *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*. ACM. 247–257.

[12] Atkinson, K. E. (1989). *An Introduction to Numerical Analysis*. John Wiley & Sons.

[13] Attneave, F. (1959). *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results*. Holt, Rinehart and Winston.

[14] Balakrishnan, S., M. J. Wainwright, and B. Yu (2017). "Statistical guarantees for the EM algorithm: From population to sample-based analysis". *The Annals of Statistics*. 45(1): 77–120.

[15] Bandeira, A. S., P. Rigollet, and J. Weed (2017). "Optimal rates of estimation for multi-reference alignment". arXiv: 1702.08546.

[16] Bar-Yossef, Z., T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan (2002). "Counting distinct elements in a data stream". In: *Proceedings of the 6th Randomization and Approximation Techniques in Computer Science*. Springer. 1–10.

[17] Bar-Yossef, Z., R. Kumar, and D. Sivakumar (2001). "Sampling algorithms: Lower bounds and applications". In: *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*. ACM. 266–275.

[18] Basharin, G. (1959). "On a statistical estimate for the entropy of a sequence of independent random variables". *Theory of Probability & Its Applications*. 4(3): 333–336.

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

[20] Benevenuto, F., T. Rodrigues, M. Cha, and V. Almeida (2009). "Characterizing user behavior in online social networks". In: *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. Association for Computing Machinery. 49–62.

[21] Berkson, J. (1980). "Minimum chi-square, not maximum likelihood! (with discussion)". *The Annals of Statistics*. 8(3): 457–487.

[22] Berry, M. J., D. K. Warland, and M. Meister (1997). "The structure and precision of retinal spike trains". *Proceedings of the National Academy of Sciences*. 94(10): 5411–5416.

[23] Bhat, S. and R. Sproat (2009). "Knowing the unseen: Estimating vocabulary size over unseen samples". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. Association for Computational Linguistics. 109–117.

[24] Birgé, L. (1983). "Approximation dans les espaces métriques et théorie de l'estimation". *Z. für Wahrscheinlichkeitstheorie und Verw. Geb.* 65(2): 181–237.

[25] Braess, D., J. Forster, T. Sauer, and H. U. Simon (2002). "How to achieve minimax expected Kullback-Leibler distance from an unknown finite distribution". In: *Algorithmic Learning Theory*. Springer. 380–394.

[26] Braess, D. and T. Sauer (2004). "Bernstein polynomials and learning theory". *Journal of Approximation Theory*. 128(2): 187–206.

[27] Bresler, G. (2015). "Efficiently learning Ising models on arbitrary graphs". In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. ACM. 771–782.

[28] Brubaker, S. C. and S. Vempala (2008). "Isotropic PCA and affine-invariant clustering". In: *IEEE 49th Annual IEEE Symposium on Foundations of Computer Science, 2008*. IEEE. 551–560.

[29] Buldygin, V. V. and Y. V. Kozachenko (1980). "Sub-Gaussian random variables". *Ukrainian Mathematical Journal*. 32(6): 483–489.

[30] Bunge, J. and M. Fitzpatrick (1993). "Estimating the number of species: A review". *Journal of the American Statistical Association*. 88(421): 364–373.

[31] Burnham, K. P. and W. S. Overton (1979). "Robust estimation of population size when capture probabilities vary among animals". *Ecology*. 60(5): 927–936.

[32] Bustamante, J. (2011). *Algebraic Approximation: A Guide to Past and Current Solutions*. Springer.

[33] Cai, T. T. and M. G. Low (2005). "Nonquadratic estimators of a quadratic functional". *The Annals of Statistics*. 33(6): 2930–2956.

[34] Cai, T. and M. G. Low (2011). "Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional". *The Annals of Statistics*. 39(2): 1012–1041.

[35] Chao, A. (1984). "Nonparametric estimation of the number of classes in a population". *Scandinavian Journal of Statistics*. 11(4): 265–270.

[36] Chao, A. and S.-M. Lee (1992). "Estimating the number of classes via sample coverage". *Journal of the American Statistical Association*. 87(417): 210–217.

[37] Charikar, M., S. Chaudhuri, R. Motwani, and V. Narasayya (2000). "Towards estimation error guarantees for distinct values". In: *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*. ACM. 268–279.

[38] Chaussé, P. (2010). “Computing generalized method of moments and generalized empirical likelihood with R”. *Journal of Statistical Software*. 34(11): 1–35. URL: http://www.jstatsoft.org/v34/i11/.

[39] Chen, J. (1995). “Optimal rate of convergence for finite mixture models”. *The Annals of Statistics*. 23(1): 221–233.

[40] Chow, C. and C. Liu (1968). “Approximating discrete probability distributions with dependence trees”. *IEEE Trans. Inf. Theory*. 14(3): 462–467.

[41] Csiszár, I. and J. Körner (1982). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press.

[42] Curto, R. E. and L. A. Fialkow (1991). “Recursiveness, positivity, and truncated moment problems”. *Houston Journal of Mathematics*. 17(4): 603–635.

[43] Darroch, J. and D. Ratcliff (1980). “A note on capture-recapture estimation”. *Biometrics*. 36: 149–153.

[44] Dasgupta, S. (1999). “Learning mixtures of Gaussians”. In: *40th Annual Symposium on Foundations of Computer Science, 1999*. IEEE. 634–644.

[45] Daskalakis, C., C. Tzamos, and M. Zampetakis (2017). “Ten steps of EM suffice for mixtures of two Gaussians”. In: *Proceedings of the 30th Annual Conference on Learning Theory (COLT 2017)*. PMLR. 704–710.

[46] Davis, P. J. (1975). *Interpolation and Approximation*. Courier Corporation.

[47] de Boor, C. (2005). “Divided differences”. *Surveys in Approximation Theory*. 1: 46–49.

[48] Deely, J. and R. Kruse (1968). “Construction of sequences estimating the mixing distribution”. *The Annals of Mathematical Statistics*. 39(1): 286–288.

[49] Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). “Maximum likelihood from incomplete data via the EM algorithm”. *Journal of the Royal Statistical Society: Series B (Methodological)*. 39(1): 1–22.

[50] DeVore, R. A. and G. G. Lorentz (1993). *Constructive Approximation*. Springer.

[51] Diaconis, P. (1987). “Application of the method of moments in probability and statistics”. In: *Moments in Mathematics*. Vol. 37. Providence, RI: Amer. Math. Soc. 125–139.

[52] Diamond, H. G. and A. Straub (2016). “Bounds for the logarithm of the Euler gamma function and its derivatives”. *Journal of Mathematical Analysis and Applications*. 433(2): 1072–1083.

[53] Ditzian, Z. and V. Totik (2012). *Moduli of Smoothness*. Vol. 9. Springer.

[54] Dobrushin, R. (1958). “A statistical problem arising in the theory of detection of signals in the presence of noise in a multi-channel system and leading to stable distribution laws”. *Theory of Probability & Its Applications*. 3(2): 161–173.

[55] Donoho, D. L. and R. C. Liu (1991). “Geometrizing rates of convergence, II”. *The Annals of Statistics*. 19: 668–701.

[56] Doss, N., Y. Wu, P. Yang, and H. H. Zhou (2020). “Optimal estimation of high-dimensional Gaussian mixtures”. arXiv: 2002.05818.

[57] Dzyadyk, V. K. and I. A. Shevchuk (2008). *Theory of Uniform Approximation of Functions by Polynomials*. Walter de Gruyter.

[58] Edelman, D. (1988). “Estimation of the mixing distribution for a normal mean with applications to the compound decision problem”. *The Annals of Statistics*. 16(4): 1609–1622.

[59] Efron, B. and R. Thisted (1976). “Estimating the number of unseen species: How many words did Shakespeare know?” *Biometrika*. 63(3): 435–447.

[60] Efron, B. (1982). “Maximum likelihood and decision theory”. English. *The Annals of Statistics*. 10(2): 340–356.

[61] Esty, W. W. (1986). “Estimation of the size of a coinage: A survey and comparison of methods”. *The Numismatic Chronicle (1966–)*. 146: 185–215.

[62] Fisher, R. A., A. S. Corbet, and C. B. Williams (1943). “The relation between the number of species and the number of individuals in a random sample of an animal population”. *Journal of Animal Ecology*. 12(1): 42–58.

[63] Freud, G. (1971). *Orthogonal Polynomials*. Pergamon.

[64] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

[65] Gao, Y., I. Kontoyiannis, and E. Bienenstock (2006). "From the entropy to the statistical structure of spike trains". In: *2006 IEEE International Symposium on Information Theory*. 645–649.

[66] Gautschi, W. (2004). *Orthogonal Polynomials: Computation and Approximation*. Oxford University Press.

[67] Genovese, C. R. and L. Wasserman (2000). "Rates of convergence for the Gaussian mixture sieve". *Annals of Statistics*. 28(4): 1105–1127.

[68] Ghosal, S. and A. W. van der Vaart (2001). "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities". *The Annals of Statistics*. 29(5): 1233–1263.

[69] Gibbs, A. L. and F. E. Su (2002). "On choosing and bounding probability metrics". *International Statistical Review*. 70(3): 419–435.

[70] "Global language monitor. Number of words in the English language" (n.d.). URL: https://www.languagemonitor.com/global-english/no-of-words/. Accessed: 2016-02-16.

[71] Golub, G. H. and J. H. Welsch (1969). "Calculation of Gauss quadrature rules". *Mathematics of Computation*. 23(106): 221–230.

[72] Good, I. and G. Toulmin (1956). "The number of new species, and the increase in population coverage, when a sample is increased". *Biometrika*. 43(1–2): 45–63.

[73] Good, I. J. (1953). "The population frequencies of species and the estimation of population parameters". *Biometrika*. 40(3–4): 237–264.

[74] Gotelli, N. J. and R. K. Colwell (2011). "Estimating species richness". *Biological Diversity: Frontiers in Measurement and Assessment*. 12: 39–54.

[75] Gradshteyn, I. S. and I. M. Ryzhik (2007). *Table of Integrals Series and Products*. Seventh edition. New York, NY: Academic.

[76] Hall, A. R. (2005). *Generalized Method of Moments*. Oxford University Press.

[77] Han, Y., J. Jiao, C.-Z. Lee, T. Weissman, Y. Wu, and T. Yu (2018a). "Entropy rate estimation for Markov chains with large state space". In: *Proceedings of the Thirty-second Conference on Neural Information Processing Systems.* Curran Associates, Inc. 9781–9792.

[78] Han, Y., J. Jiao, and T. Weissman (2015a). "Adaptive estimation of Shannon entropy". In: *2015 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 1372–1376.

[79] Han, Y., J. Jiao, and T. Weissman (2015b). "Does Dirichlet prior smoothing solve the Shannon entropy estimation problem?" In: *2015 IEEE International Symposium on Information Theory (ISIT).* IEEE. 1367–1371.

[80] Han, Y., J. Jiao, and T. Weissman (2018b). "Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance". In: *Proceedings of the 31st Conference on Learning Theory.* PMLR. 3189–3221.

[81] Han, Y., J. Jiao, T. Weissman, and Y. Wu (2017). "Optimal rates of entropy estimation over Lipschitz balls". *To appear in Annals of Statistics.* arXiv: 1711.02141.

[82] Hansen, L. P. (1982). "Large sample properties of generalized method of moments estimators". *Econometrica.* 50(4): 1029–1054.

[83] Härdle, W., G. Kerkyacharian, D. Picard, and A. Tsybakov (2012). *Wavelets, Approximation, and Statistical Applications.* Vol. 129. Springer.

[84] Hardt, M. and E. Price (2015). "Tight bounds for learning a mixture of two gaussians". In: *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing.* ACM. 753–760.

[85] Harris, B. (1975). "The statistical estimation of entropy in the non-parametric case". In: *Topics in Information Theory.* Ed. by I. Csiszár and P. Elias. Vol. 16. Springer Netherlands. 323–355.

[86] Heinrich, P. and J. Kahn (2018). "Strong identifiability and optimal minimax rates for finite mixture estimation". *The Annals of Statistics.* 46(6A): 2844–2870.

[87] Hopkins, S. B. and J. Li (2018). "Mixture models, robustness, and sum of squares proofs". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM. 1021–1034.

[88] Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis*. 2nd ed. Cambridge University Press.

[89] Hou, W.-C., G. Ozsoyoglu, and B. K. Taneja (1988). "Statistical estimators for relational algebra expressions". In: *Proceedings of the Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM. 276–287.

[90] Hsu, D. and S. M. Kakade (2013). "Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions". In: *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*. ACM. 11–20.

[91] Huang, S.-P. and B. Weir (2001). "Estimating the total number of alleles using a sample coverage method". *Genetics*. 159(3): 1365–1373.

[92] Ibragimov, I. (2001). "Estimation of analytic functions". *Lecture Notes-Monograph Series*. 36: 359–383, Institute of Mathematical Statistics.

[93] Ibragimov, I. and R. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory*. Springer.

[94] Ibragimov, I., A. Nemirovskii, and R. Khas'minskii (1987). "Some problems on nonparametric estimation in Gaussian white noise". *Theory of Probability & Its Applications*. 31(3): 391–406.

[95] Ionita-Laza, I., C. Lange, and N. M. Laird (2009). "Estimating the number of unseen variants in the human genome". *Proceedings of the National Academy of Sciences*. 106(13): 5008–5013.

[96] Ismail, M. E. H. (2005). *Classical and Quantum Orthogonal Polynomials in One Variable*. Vol. 13. Cambridge University Press.

[97] Ivanov, K. G. (1983). "On a new characteristic of functions. II. Direct and converse theorems for the best algebraic approximation in $C[-1,1]$ and $L_p[-1,1]$". *Pliska*. 5: 151–163.

[98] Jewell, N. P. (1982). "Mixtures of exponential distributions". *The Annals of Statistics*. 10(2): 479–484.

[99] Jiao, J., Y. Han, and T. Weissman (2016). “Beyond maximum likelihood: Boosting the Chow-Liu algorithm for large alphabets”. In: *2016 50th Asilomar Conference on Signals, Systems and Computers*. IEEE. 321–325.

[100] Jiao, J., H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman (2013). “Universal estimation of directed information”. *IEEE Trans. Inf. Theory*. 59(10): 6220–6242.

[101] Jiao, J., K. Venkat, Y. Han, and T. Weissman (2015). “Minimax estimation of functionals of discrete distributions”. *IEEE Transactions on Information Theory*. 61(5): 2835–2885.

[102] Jiao, J., K. Venkat, Y. Han, and T. Weissman (2017). “Maximum likelihood estimation of functionals of discrete distributions”. *IEEE Transactions on Information Theory*. 63(10): 6774–6798.

[103] Juditsky, A. B. and A. S. Nemirovski (2009). “Nonparametric estimation by convex programming”. *The Annals of Statistics*. 37(5A): 2278–2300.

[104] Kalai, A. T., A. Moitra, and G. Valiant (2010). “Efficiently learning mixtures of two Gaussians”. In: *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*. ACM. 553–562.

[105] Kannan, R., H. Salmasian, and S. Vempala (2005). “The spectral method for general mixture models”. In: *International Conference on Computational Learning Theory*. Springer. 444–457.

[106] Karlin, S. and L. S. Shapley (1953). *Geometry of Moment Spaces*. No. 12. American Mathematical Society.

[107] Karlis, D. and E. Xekalaki (2003). “Choosing initial values for the EM algorithm for finite mixtures”. *Computational Statistics & Data Analysis*. 41(3): 577–590.

[108] Karlis, D. and E. Xekalaki (2005). “Mixed Poisson distributions”. *International Statistical Review*. 73(1): 35–58.

[109] Kelly, B., A. Wagner, T. Tularak, and P. Viswanath (2013). “Classification of homogeneous data with large alphabets”. *IEEE Transactions on Information Theory*. 59(2): 782–795.

[110] Kiefer, J. and J. Wolfowitz (1956). “Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters”. *The Annals of Mathematical Statistics*: 887–906.

[111] Kim, A. K. (2014). “Minimax bounds for estimation of normal mixtures”. *Bernoulli*. 20(4): 1802–1818.

[112] Knudson, K. C. and J. W. Pillow (2013). “Spike train entropy-rate estimation using hierarchical Dirichlet process priors”. In: *Proceedings of the Twenty-seventh Conference on Neural Information Processing Systems*. Curran Associates, Inc. 2076–2084.

[113] Koenker, R. and I. Mizera (2014). “Convex optimization, shape constraints, compound decisions, and empirical Bayes rules”. *Journal of the American Statistical Association*. 109(506): 674–685.

[114] Kong, W. and G. Valiant (2017). “Spectrum estimation from samples”. *The Annals of Statistics*. 45(5): 2218–2247.

[115] Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.

[116] Krawtchouk, M. (1932). “Sur le problème de moments”. In: *ICM Proceedings*. Available at https://www.mathunion.org/fileadmin/ICM/Proceedings/ICM1932.2/ICM1932.2.ocr.pdf. 127–128.

[117] Krein, M. G. and A. A. Nudel’man (1977). *The Markov Moment Problem and Extremal Problems*. American Mathematical Society.

[118] Kybic, J. (2004). “High-dimensional mutual information estimation for image registration”. In: *2004 International Conference on Image Processing, 2004. ICIP’04*. Vol. 3. IEEE. 1779–1782.

[119] Laird, N. (1978). “Nonparametric maximum likelihood estimation of a mixing distribution”. *Journal of the American Statistical Association*. 73(364): 805–811.

[120] Lasserre, J. B. (2009). *Moments, Positive Polynomials and Their Applications*. Vol. 1. World Scientific.

[121] Laurent, B. (1996). “Efficient estimation of integral functionals of a density”. *The Annals of Statistics*. 24(2): 659–681.

[122] Le Cam, L. (1973). “Convergence of estimates under dimensionality restrictions”. *The Annals of Statistics*. 1(1): 38–53.

[123] Lepski, O., A. Nemirovski, and V. Spokoiny (1999). “On estimation of the L_r norm of a regression function”. *Probability Theory and Related Fields*. 113(2): 221–253.

[124] Li, J. and L. Schmidt (2017). “Robust and proper learning for mixtures of Gaussians via systems of polynomial inequalities”. In: *Proceedings of the 30th Annual Conference on Learning Theory (COLT 2017).* PMLR. 1302–1382.

[125] Lindsay, B. G. (1981). “Properties of the maximum likelihood estimator of a mixing distribution”. In: *Statistical Distributions in Scientific Work*. Springer. 95–109.

[126] Lindsay, B. G. (1989). “Moment matrices: Applications in mixtures”. *The Annals of Statistics*. 17(2): 722–740.

[127] Lindsay, B. G. (1995). “Mixture models: Theory, geometry and applications”. In: *NSF-CBMS Regional Conference Series in Probability and Statistics*. 5: I–163, American Statistical Association.

[128] Lo, S.-H. (1992). “From the species problem to a general coverage problem via a new interpretation”. *The Annals of Statistics*. 20(2): 1094–1109.

[129] Lu, Y. and H. H. Zhou (2016). “Statistical and computational guarantees of Lloyd’s algorithm and its variants”. arXiv: 1612.02099.

[130] Mainen, Z. F. and T. J. Sejnowski (1995). “Reliability of spike timing in neocortical neurons”. *Science*. 268(5216): 1503–1506.

[131] Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

[132] McNeil, D. R. (1973). “Estimating an author’s vocabulary”. *Journal of the American Statistical Association*. 68(341): 92–96.

[133] Meng, X., J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, *et al.* (2016). “Mllib: Machine learning in apache spark”. *The Journal of Machine Learning Research*. 17(1): 1235–1241.

[134] Meng, X.-L. and D. Van Dyk (1997). “The EM algorithm—An old folk-song sung to a fast new tune”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 59(3): 511–567.

[135] Miller, G. A. (1955). “Note on the bias of information estimates”. *Information Theory in Psychology: Problems and Methods*. 2: 95–100.

[136] Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.

[137] Mitzenmacher, M. and E. Upfal (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press.

[138] Moitra, A. and G. Valiant (2010). “Settling the polynomial learnability of mixtures of Gaussians”. In: *2010 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS).* IEEE. 93–102.

[139] Morris, C. N. (1982). “Natural exponential families with quadratic variance functions”. *The Annals of Statistics.* 10(1): 65–80.

[140] Naughton, J. F. and S. Seshadri (1990). “On estimating the size of projections”. In: *International Conference on Database Theory.* Springer. 499–513.

[141] Nemenman, I., W. Bialek, and R. R. de Ruyter van Steveninck (2004). “Entropy and information in neural spike trains: Progress on the sampling problem”. *Physical Review E.* 69(5): 056111.

[142] Nemirovski, A. (2003). “On tractable approximations of randomly perturbed convext constaints”. In: *Proceedings of the 42nd IEEE Conference on Decision and Control.* IEEE. 2419–2422.

[143] Nikolsky, S. (1946). “Approximation of functions in the mean by trigonometrical polynomials”. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya.* 10(3): 207–256.

[144] Oppenheim, A. V. (1999). *Discrete-Time Signal Processing.* Pearson Education India.

[145] Orlitsky, A., N. P. Santhanam, and J. Zhang (2004). “Universal compression of memoryless sources over unknown alphabets”. *IEEE Transactions on Information Theory.* 50(7): 1469–1481.

[146] Orlitsky, A. and A. T. Suresh (2015). “Competitive distribution estimation: Why is good-turing good”. In: *Proceedings of the Twenty-Ninth Conference on Neural Information Processing Systems.* Curran Associates, Inc. 2143–2151.

[147] Orlitsky, A., A. T. Suresh, and Y. Wu (2016). “Optimal prediction of the number of unseen species”. *Proceedings of the National Academy of Sciences (PNAS).* 113(47): 13283–13288.

[148] “Oxford English Dictionary” (n.d.). http://public.oed.com/about/. Accessed: 2016-02-16.

[149] Paninski, L. (2003). "Estimation of entropy and mutual information". *Neural Computation*. 15(6): 1191–1253.

[150] Paninski, L. (2004). "Estimating entropy on m bins given fewer than m samples". *IEEE Transactions on Information Theory*. 50(9): 2200–2203.

[151] Pearson, K. (1894). "Contributions to the mathematical theory of evolution". *Philosophical Transactions of the Royal Society of London. A*. 185: 71–110.

[152] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine learning in python". *Journal of Machine Learning Research*. 12: 2825–2830.

[153] Petrushev, P. P. and V. A. Popov (2011). *Rational Approximation of Real Functions*. Cambridge University Press.

[154] Pilla, R. S. and B. G. Lindsay (2001). "Alternative EM methods for nonparametric finite mixture models". *Biometrika*. 88(2): 535–550.

[155] Pinsker, M. S. (1980). "Optimal filtering of square-integrable signals in Gaussian noise". *Problemy Peredachi Informatsii*. 16(2): 52–68.

[156] Plotkin, N. T. and A. J. Wyner (1996). "An entropy estimator algorithm and telecommunications applications". In: *Maximum Entropy and Bayesian Methods*. Vol. 62. *Fundamental Theories of Physics*. Springer Netherlands. 351–363.

[157] Pluim, J. P., J. A. Maintz, and M. A. Viergever (2003). "Mutual-information-based registration of medical images: A survey". *IEEE Transactions on Medical Imaging*. 22(8): 986–1004.

[158] Polyanskiy, Y., A. T. Suresh, and Y. Wu (2017). "Sample complexity of population recovery". In: *Proceedings of Conference on Learning Theory (COLT)*. Amsterdam, Netherland. arXiv: 1702.05574.

[159] Polyanskiy, Y. and Y. Wu (2019). "Dualizing Le Cam's method, with applications to estimating the unseens". arXiv: 1902.05616.

[160] Porta, A., S. Guzzetti, N. Montano, R. Furlan, M. Pagani, A. Malliani, and S. Cerutti (2001). "Entropy, entropy rate, and pattern classification as tools to typify complexity in short heart period variability series". *IEEE Transactions on Biomedical Engineering*. 48(11): 1282–1291.

[161] Portilla, J., V. Strela, M. J. Wainwright, and E. P. Simoncelli (2003). "Image denoising using scale mixtures of Gaussians in the wavelet domain". *IEEE Transactions on Image Processing*. 12(11): 1338–1351.

[162] Prasolov, V. V. (2009). *Polynomials*. Vol. 11. Springer.

[163] Quinn, C. J., N. Kiyavash, and T. P. Coleman (2013). "Efficient methods to compute optimal tree approximations of directed information graphs". *IEEE Trans. Signal Process.* 61(12): 3173–3182.

[164] Rao, B. P. (2014). *Nonparametric Functional Estimation*. Academic Press.

[165] Raskhodnikova, S., D. Ron, A. Shpilka, and A. Smith (2009). "Strong lower bounds for approximating distribution support size and the distinct elements problem". *SIAM Journal on Computing*. 39(3): 813–842.

[166] Redner, R. A. and H. F. Walker (1984). "Mixture densities, maximum likelihood and the EM algorithm". *SIAM Review*. 26(2): 195–239.

[167] Reimer, M. (2012). *Multivariate Polynomial Approximation*. Vol. 144. Birkhäuser.

[168] Rieke, F., W. Bialek, D. Warland, and R. d. R. van Steveninck (1999). *Spikes: Exploring the Neural Code*. The MIT Press.

[169] Rivlin, T. J. (1981). *An Introduction to the Approximation of Functions*. Dover.

[170] Rockafellar, R. T. (1974). *Conjugate Duality and Optimization*. Vol. 16. Siam.

[171] Rudin, W. (2006). *Real and Complex Analysis*. Tata McGraw-Hill Education.

[172] Ruyter van Steveninck, R. R. de, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek (1997). "Reproducibility and variability in neural spike trains". *Science*. 275(5307): 1805–1808.

[173] Saha, S., A. Guntuboyina, *et al.* (2020). “On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising”. *Annals of Statistics*. 48(2): 738–762.

[174] Schmüdgen, K. (2017). *The Moment Problem*. Springer.

[175] Seidel, W., K. Mosler, and M. Alker (2000). “A cautionary note on likelihood ratio tests in mixture models”. *Annals of the Institute of Statistical Mathematics*. 52(3): 481–487.

[176] Shannon, C. E. (1948). “A mathematical theory of communication”. *Bell System Technical Journal*. 27: 379–423, 623–656.

[177] Shohat, J. A. and J. D. Tamarkin (1943). *The Problem of Moments*. No. 1. American Mathematical Society.

[178] Steele, J. M. (1986). “An Efron-Stein inequality for nonsymmetric statistics”. *The Annals of Statistics*. 14(2): 753–758.

[179] Stein, C. (1986). “Lectures on the theory of estimation of many parameters”. *Journal of Soviet Mathematics*. 34(1): 1373–1403. [Zap. Nauchn. Sem. LOMI, 1977, Volume 74, Pages 4–65].

[180] Stoer, J. and R. Bulirsch (2002). *Introduction to Numerical Analysis*. 3rd edition. New York, NY: Springer.

[181] Stone, C. J. (1980). “Optimal rates of convergence for nonparametric estimators”. *The Annals of Statistics*. 8(6): 1348–1360.

[182] Strasser, H. (1985). *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. Berlin, Germany: Walter de Gruyter.

[183] Strong, S. P., R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek (1998). “Entropy and information in neural spike trains”. *Phys. Rev. Lett.* 80(1): 197–200.

[184] Szegö, G. (1975). *Orthogonal Polynomials*. 4th. Providence, RI: American Mathematical Society.

[185] Thisted, R. and B. Efron (1987). “Did Shakespeare write a newly-discovered poem?” *Biometrika*. 74(3): 445–455.

[186] Timan, A. F. (1963). *Theory of Approximation of Functions of a Real Variable*. Pergamon Press.

[187] Tsai, A., J. W. Fisher, C. Wible, W. M. Wells, J. Kim, and A. S. Willsky (1999). "Analysis of functional MRI data using mutual information". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 473–480.

[188] Tsai, A., W. Wells, C. Tempany, E. Grimson, and A. Willsky (2004). "Mutual information in coupled multi-shape model for medical image segmentation". *Medical Image Analysis*. 8(4): 429–445.

[189] Tse, D. and P. Viswanath (2005). *Fundamentals of Wireless Communication*. Cambridge University Press.

[190] Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. New York, NY: Springer.

[191] Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill.

[192] Valiant, G. (2017). Private communication.

[193] Valiant, G. and P. Valiant (2010). "A CLT and tight lower bounds for estimating entropy". *Electronic Colloquium on Computational Complexity (ECCC)*. 17(179): 1–32.

[194] Valiant, G. and P. Valiant (2011a). "Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs". In: *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*. ACM. 685–694.

[195] Valiant, G. and P. Valiant (2011b). "The power of linear estimators". In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 403–412.

[196] Valiant, P. (2008). "Testing symmetric properties of distributions". In: *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing. STOC '08*. ACM. 383–392.

[197] Valiant, P. (2011). "Testing symmetric properties of distributions". *SIAM Journal on Computing*. 40(6): 1927–1968.

[198] Valiant, P. and G. Valiant (2013). "Estimating the unseen: Improved estimators for entropy and other properties". In: *Proceedings of the Twenty-Seventh Conference on Neural Information Processing Systems*. Curran Associates, Inc. 2157–2165.

[199] Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

[200] Vempala, S. and G. Wang (2004). "A spectral algorithm for learning mixture models". *Journal of Computer and System Sciences*. 68(4): 841–860.

[201] Villani, C. (2003). *Topics in Optimal Transportation*. Providence, RI: American Mathematical Society.

[202] Villani, C. (2008). *Optimal Transport: Old and New*. Berlin: Springer.

[203] Vinck, M., F. P. Battaglia, V. B. Balakirsky, A. H. Vinck, and C. M. Pennartz (2012). "Estimation of the entropy based on its polynomial representation". *Physical Review E*. 85(5): 051139.

[204] Wagner, A. B., P. Viswanath, and S. R. Kulkarni (2011). "Probability estimation in the rare-events regime". *IEEE Trans. Inf. Theory*. 57(6): 3207–3229.

[205] Wainwright, M. J. and E. P. Simoncelli (2000). "Scale mixtures of Gaussians and the statistics of natural images". In: *Proceedings of the Thirteenth Conference on Neural Information Processing Systems (NIPS 1999)*. MIT Press. 855–861.

[206] Wolkowicz, H., R. Saigal, and L. Vandenberghe (2012). *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*. Vol. 27. Springer.

[207] Wu, Y. and S. Verdú (2010). "The impact of constellation cardinality on Gaussian channel capacity". In: *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE. 620–628.

[208] Wu, Y. and P. Yang (2016). "Minimax rates of entropy estimation on large alphabets via best polynomial approximation". *IEEE Transactions on Information Theory*. 62(6): 3702–3720.

[209] Wu, Y. and P. Yang (2018). "Sample complexity of the distinct elements problem". *Mathematical Statistics and Learning*. 1(1): 37–72.

[210] Wu, Y. and P. Yang (2019). "Chebyshev polynomials, moment matching, and optimal estimation of the unseen". *The Annals of Statistics*. 47(2): 857–883.

[211] Wu, Y. and P. Yang (2020a). "Optimal estimation of Gaussian mixtures via denoised method of moments". *The Annals of Statistics*. 48(4): 1981–2007.

[212] Wu, Y. and P. Yang (2020b). "Supplement to 'optimal estimation of Gaussian mixtures via denoised method of moments'". DOI: 10.1214/19-AOS1873SUPP.

[213] Xu, J., D. J. Hsu, and A. Maleki (2016). "Global analysis of expectation maximization for mixtures of two Gaussians". In: *Proceedings of the Thirtieth Conference on Neural Information Processing Systems*. Curran Associates, Inc. 2676–2684.

[214] Xu, L. and M. I. Jordan (1996). "On convergence properties of the EM algorithm for Gaussian mixtures". *Neural Computation*. 8(1): 129–151.

[215] Yang, P. (2016). "Optimal property estimation on large alphabets: Fundamental limits and fast algorithms". *MA thesis*. University of Illinois at Urbana-Champaign.

[216] Yang, Y. and A. R. Barron (1999). "Information-theoretic determination of minimax rates of convergence". *The Annals of Statistics*. 27(5): 1564–1599.

[217] Yu, B. (1997). "Assouad, Fano, and Le Cam". *Festschrift for Lucien Le Cam*: 423–435, Springer, New York.

[218] Zhang, C.-H. (2009). "Generalized maximum likelihood estimation of normal mixture densities". *Statistica Sinica*. 19: 1297–1318.

[219] Zou, J., G. Valiant, P. Valiant, K. Karczewski, S. O. Chan, K. Samocha, M. Lek, S. Sunyaev, M. Daly, and D. G. MacArthur (2016). "Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects". *Nature Communications*. 7: 13293.