Math. Z. 247, 183–199 (2004) **Mathematische Zeitschrift**

DOI: 10.1007/s00209-003-0633-9

The Markov moment problem and de Finetti's theorem: Part I

Persi Diaconis[1], **David Freedman**[2]

[1] Department of Mathematics and of Statistics, Stanford University, Stanford, CA 94305, USA

[2] Department of Statistics and of Mathematics, University of California, Berkeley, CA 94720-3860, USA (e-mail: freedman@stat.berkeley.edu)

Received: 1 February 2003; in final form: 18 June 2003 /
Published online: 14 January 2004 –

Dedicated to the memory of Sergei Kerov

Abstract The Markov moment problem is to characterize sequences $s_0, s_1, s_2, \ldots$ admitting the representation $s_n = \int_0^1 x^n f(x)\,dx$, where $f(x)$ is a probability density on $[0, 1]$ and $0 \le f(x) \le c$ for almost all x. There are well-known characterizations through complex systems of non-linear inequalities on $\{s_n\}$. Necessary and sufficient linear conditions are the following: $s_0 = 1$, and

$$0 \le (-1)^{n-j}\binom{n}{j}\Delta^{n-j}s_j \le c/(n+1)$$

for all $n = 0, 1, \ldots$ and $j = 0, 1, \ldots, n$. Here, Δ is the forward difference operator. This result is due to Hausdorff. We give a new proof with some ancillary results, for example, characterizing monotone densities. Then we make the connection to de Finetti's theorem, with characterizations of the mixing measure.

Introduction

We begin by reviewing the Hausdorff moment problem. Then we take up the Markov moment problem, with a solution due to Hausdorff (1923). Although this work was discussed in an earlier generation of texts (Shohat and Tamarkin, 1943, pp. 98–101; Widder, 1946, pp. 109–12; Hardy, 1949, pp. 272–3), it seems less well known today than the one due to the Russian school. Next, we sketch some generalizations and the connection to de Finetti's theorem. We close with some historical notes, including a brief statement of the Russian work. We believe that our Theorem 4 is new, along with the local theorems, the applications to Bayesian statistics (Theorems 8 and 9), and the characterization of measures with monotone densities

(Theorem 10). Many of the results in this paper can be seen as answers to one facet or another of the following question: what can you learn about a measure from the moments, and how is it to be done?

The Hausdorff moment problem

Let $s_0, s_1, s_2, \ldots$ be a sequence of real numbers. When is there a probability measure μ on the unit interval such that s_n is the nth moment of μ? In other words, we seek the necessary and sufficient conditions on $\{s_n\}$ for there to exist a probability μ with

$$s_n = \int_0^1 x^n \, \mu(dx) \quad \text{for} \quad n = 0, 1, \ldots.$$

This is the Hausdorff moment problem.

To state Hausdorff's solution, let $\Delta t_n = t_{n+1} - t_n$ be the forward difference operator. Define an auxiliary sequence as

$$s_{n,j} = (-1)^{n-j} \binom{n}{j} \Delta^{n-j} s_j \tag{1}$$

for $n = 0, 1, \ldots$ and $j = 0, 1, \ldots, n$. By convention, $\Delta^0 s_j = s_j$. Thus,

$$\begin{aligned} s_{j,j} &= s_j, \\ s_{j+1,j} &= (j+1)(s_j - s_{j+1}), \\ s_{j+2,j} &= \tfrac{1}{2}(j+1)(j+2)(s_{j+2} - 2s_{j+1} + s_j), \end{aligned}$$

and so forth. The reason for introducing the binomial coefficients will be discussed later.

Theorem 1. *Given a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1). *There exists a probability measure μ on* $[0, 1]$ *such that $\{s_n\}$ is the moment sequence of μ if and only if $s_0 = 1$, and $0 \le s_{n,j}$ for all n and j. Then μ is unique.*

This theorem is due to Hausdorff (1921), but Feller (1971, pp. 224–28) may be more accessible; the proof will not be repeated here. The "Hausdorff moment condition" is that $0 \le s_{n,j}$ for all n and j.

The Markov moment problem

The "Markov moment problem" is to characterize moments of probabilities that have uniformly bounded densities, which constrains μ in Theorem 1 to have the form $\mu(dx) = f(x)\,dx$, where $f \le c$ a.e. Of course, $f \ge 0$ a.e. and $\int_0^1 f\,dx = 1$, so $c \ge 1$. Hausdorff's solution is presented as Theorem 2.

Theorem 2. *Given a positive real number c, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1). *There exists a probability measure μ on* $[0, 1]$ *such that*

(i) $\{s_n\}$ *is the moment sequence of* μ, *and*
(ii) μ *is absolutely continuous, and*
(iii) $d\mu/dx$ *is almost everywhere bounded above by* c,

if and only if $s_0 = 1$, *and* $0 \le s_{n,j} \le c/(n+1)$ *for all* n *and* j. *Then* μ *is unique.*

Our proof will use the following lemma.

Lemma 1. *Suppose* $\{s_n\}$ *is the moment sequence of the probability* μ *on* $[0, 1]$; *define the auxiliary sequence by* (1). *Then*

(a) $s_{n,j} = \binom{n}{j} \int_0^1 x^j (1-x)^{n-j}\, \mu(dx)$.
(b) *If* μ *is Lebesgue measure, then* $s_n = 1/(n+1)$.
(c) *If* μ *is Lebesgue measure, then* $s_{n,j} = s_{n,n} = s_n = 1/(n+1)$.

Proof. Claim (a). Induction on $n = j, j+1, \ldots$.
Claim (b). Integration.
Claim (c). This just depends on the beta integral (Feller, 1971, p. 47):

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \quad \text{for positive real } \alpha, \beta. \tag{2}$$

Remarks. (i) Property (b) characterizes Lebesgue measure, in view of the uniqueness part of Theorem 1. Likewise, $s_{n,j} = s_{n,n}$ for all n and $j = 0, \ldots, n$ is a characterization, as in (c). Indeed,

$$\sum_{j=0}^{n} \binom{n}{j} x^j (1-x)^{n-j} = 1$$

for all x in the unit interval—after all, $[x + (1-x)]^n = 1$. Lemma 1a implies

$$\sum_{j=0}^{n} s_{n,j} = 1.$$

If the $s_{n,j}$ are equal for all $j = 0, 1, \ldots, n$, each must be $1/(n+1)$, so $s_n = s_{n,n} = 1/(n+1)$ for all $n = 0, 1, \ldots$. In essence, this characterization of the uniform distribution on $[0, 1]$ is due to Bayes (1764): see Stigler (1986, pp. 128–9).

(ii) Without the binomial coefficients in (1), the upper bound on $s_{n,j}$ in Theorem 2 would be more cumbersome to state. A deeper justification may be given by formulas (1.8) and (3.7) in Feller (1971, pp. 221, 225).

(iii) The condition $s_0 = 1$ may be dropped in Theorems 1 and 2; then μ is a finite positive measure, of total mass s_0. Indeed, $\sum_{j=0}^{n} s_{n,j} = s_0$ for any sequence $\{s_n\}$; this can be proved directly, or see (1.9) in Feller (1971, p. 221).

Proof of Theorem 2. Suppose conditions (i), (ii), and (iii) hold true. The conditions on $s_{n,j}$ follow from Lemma 1. Conversely, suppose the conditions on $s_{n,j}$ hold true. Theorem 1 shows the existence (and uniqueness) of a probability measure μ whose

moment sequence is $\{s_n\}$. What remains to be seen is that μ is absolutely continuous, having a density bounded by c. If g is a non-negative continuous function on $[0, 1]$, its nth approximating Bernstein polynomial is by definition

$$B_{n,g}(x) = \sum_{j=0}^{n} g\left(\frac{j}{n}\right)\binom{n}{j}x^j(1-x)^{n-j}.$$

We claim that

$$\int_0^1 B_{n,g}(x)\,\mu(dx) \le c\int_0^1 B_{n,g}(x)\,dx.$$

Indeed, the left side is $\sum_j g(j/n)s_{n,j}$ by Lemma 1a. The right side is $\sum_j g(j/n)[c/(n+1)]$ by Lemma 1c. Finally, use the condition that $s_{n,j} \le c/(n+1)$.

Of course, $B_{n,g}$ converges to g uniformly as $n \to \infty$: see Feller (1971, pp. 222–4), or Lorentz (1966) for a more detailed discussion. So, for all non-negative continuous g,

$$\int_0^1 g(x)\,\mu(dx) \le c\int_0^1 g(x)\,dx. \tag{3}$$

Let $\mathcal{G}$ be the set of Borel measurable functions g on $[0, 1]$ with $0 \le g \le 1$. Let $\mathcal{G}_1$ consist of the $g \in \mathcal{G}$ for which (3) holds. Then $\mathcal{G}_1$ contains all the continuous functions in $\mathcal{G}$ and is closed under pointwise limits, so $\mathcal{G}_1 = \mathcal{G}$. Put $g = 1_A$, the indicator function of a Borel set A, to conclude that $\mu(A) \le c\lambda(A)$, where λ is Lebesgue measure. Now μ is absolutely continuous; denote the Radon-Nikodym derivative $d\mu/dx$ by f. Let

$$A = \{x : 0 \le x \le 1 \;\&\; f(x) > c\}.$$

If $\lambda(A) > 0$, then

$$\mu(A) = \int_A f(x)\,dx > c\lambda(A).$$

But we have already seen that $\mu(A) \le c\lambda(A)$. This contradiction shows that $\lambda(A) = 0$, proving Theorem 2. □

Example 1. Let $f(x) = 1/(2\sqrt{x})$ on $(0, 1]$. This density is unbounded, but its nth moment is $s_n = 1/(2n+1) \le 1/(n+1)$. Thus, the simple condition $s_n \le c/(n+1)$ is not sufficient to make the density bounded: auxiliary conditions are needed. For our f, $(n+1)s_{n,j}$ is unbounded. Indeed, $s_{n,j}$ can be computed explicitly, using Lemma 1a and the formula for the beta integral (2):

$$\begin{aligned} s_{n,j} &= \frac{1}{2}\binom{n}{j}\int_0^1 x^j(1-x)^{n-j}\frac{1}{\sqrt{x}}\,dx \\ &= \frac{1}{2}\binom{n}{j}\frac{\Gamma(j+\frac{1}{2})\Gamma(n-j+1)}{\Gamma(n+\frac{3}{2})} \\ &= \frac{1}{2}\frac{\Gamma(j+\frac{1}{2})}{\Gamma(j+1)}\frac{\Gamma(n+1)}{\Gamma(n+\frac{3}{2})}. \end{aligned}$$

By Stirling's formula, $\log \Gamma(x) = (x - \frac{1}{2}) \log(x + k) - x + O(1)$ as x gets large, for any constant k. Hence

$$\log(n+1) + \log \Gamma(n+1) - \log \Gamma(n + \tfrac{3}{2}) = \tfrac{1}{2} \log(n+1) + O(1).$$

So

$$\lim_{n \to \infty} (n+1) s_{n,j} = \infty$$

for any fixed j. The boundedness condition of Theorem 2 is not satisfied.

Example 2. The moments of the "Cantor measure" may be of interest in connection with Theorem 2. The Cantor measure is the distribution of $2 \sum_{j=1}^{\infty} X_j / 3^j$, the X_j being independent and identically distributed, $X_j = 0$ with probability 1/2 and $X_j = 1$ with probability 1/2. This measure is uniform on the Cantor set, and is therefore purely singular. For $n \geq 2$, the nth moment is

$$s_n > \frac{1}{2} \Big(1 - \frac{1}{n}\Big)^n \frac{1}{n^{\log_3 2}}.$$

Indeed, the Cantor measure assigns mass 2^{-m} to the interval $[1 - 3^{-m}, 1]$, so

$$s_n \geq 2^{-m} (1 - 3^{-m})^n$$

for any positive integer m. Now choose m with $\log_3 n \leq m < 1 + \log_3 n$. In particular,

$$\lim (n+1) s_n = \infty.$$

See Grabner and Prodinger (1996) for more detailed estimates.

L_p densities

Theorem 2 characterizes the moment sequences of probabilities with L_∞ densities on $[0, 1]$. The next result (also due to Hausdorff) characterizes L_p densities for $p > 1$. To state the theorem, define

$$c_n = \Big\{ \frac{1}{n+1} \sum_{j=0}^{n} \big[(n+1) s_{n,j} \big]^p \Big\}^{1/p}. \tag{4}$$

Theorem 3. *Given real numbers $p > 1$ and $0 < c < \infty$, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1)*, and c_n by* (4)*. There exists a probability measure μ on* $[0, 1]$ *such that*

(i) *$\{s_n\}$ is the moment sequence of μ, and*
(ii) *μ is absolutely continuous, and*
(iii) *$d\mu/dx$ is in L_p with p-norm at most c,*

if and only if $s_0 = 1$, and $0 \leq s_{n,j}$ for all n and j, and $c_n \leq c$.

So far, absolute continuity is defined relative to Lebesgue measure, but Lebesgue measure can be replaced by any other probability ν on $[0, 1]$. To avoid trivial complications, suppose ν assigns positive mass to the open unit interval $(0, 1)$. Let t_n be the moment sequence of ν, and $t_{n,j}$ the corresponding auxiliary sequence defined by (1) with t_n in place of s_n. Lemma 1a confirms that $t_{n,j} > 0$. Replace the definition (4) by

$$c_n = \left\{ \sum_{j=0}^{n} t_{n,j} \left(\frac{s_{n,j}}{t_{n,j}} \right)^p \right\}^{1/p}. \tag{5}$$

Theorem 4. *Let ν be a probability on $[0, 1]$, assigning positive mass to $(0, 1)$. Given real numbers $p > 1$ and $0 < c < \infty$, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation (1), and c_n by (5) rather than (4). There exists a probability measure μ on $[0, 1]$ such that*

(i) *$\{s_n\}$ is the moment sequence of μ, and*
(ii) *$\mu << \nu$, and*
(iii) *$d\mu/d\nu$ is in L_p with p-norm at most c,*

if and only if $s_0 = 1$, and $0 \le s_{n,j}$ for all n and j, and $c_n \le c$.

In (iii), the p-norm of $d\mu/d\nu$ is relative to ν, i.e., $\left(\int (d\mu/d\nu)^p \, d\nu \right)^{1/p}$. Theorem 3 is a special case of Theorem 4; our proof of the latter depends on the connection with de Finetti's theorem, which is explained next. Let $X_1, X_2, \ldots$ be random variables taking only the values 0 and 1. The sequence is "exchangeable" if the joint distribution is invariant under finite permutations, for example,

$$P\{X_1 = 1, X_2 = 0, X_3 = 1\} = P\{X_1 = 0, X_2 = 1, X_3 = 1\}.$$

Either the random variables can be permuted, or the values.

Theorem 5. *Let $e_1, e_2, \ldots$ be 0 or 1. The 0–1 valued random variables $X_1, X_2, \ldots$ are exchangeable if and only if there is a probability measure μ on $[0, 1]$ such that*

$$P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \int_0^1 \theta^{\Sigma e_i} (1 - \theta)^{n - \Sigma e_i} \, \mu(d\theta), \tag{6}$$

for all n and e_i. Then μ is unique.

This theorem is due to de Finetti (1931, 1937); for a review, see Hewitt and Savage (1955). The "if" part is straightforward. Necessity is more subtle because μ must be constructed, but Hausdorff's theorem can be used (Feller, 1971, pp. 228–9). The proof of Theorem 5 will not be detailed here. Before applying the theorem, we explain how the auxiliary sequence (1) connects to (6). Suppose the X_i are exchangeable, and $S_n = X_1 + X_2 + \cdots + X_n$. Let s_n be the moment sequence of μ in Theorem 5, and define $s_{n,j}$ by (1). Fix n and j with $0 \le j \le n$. Fix some particular finite sequence $e_1, e_2, \ldots, e_n$ of 0s and 1s whose sum is j. Then

$$P\{S_n = j\} = \binom{n}{j} P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \binom{n}{j} \int_0^1 x^j (1 - x)^{n-j} \, \mu(dx).$$

By Lemma 1a,

$$P\{S_n = j\} = s_{n,j}. \tag{7}$$

The notation is flawed, in that s_n is a moment of μ rather than a value of S_n.

Proof of Theorem 4. If $s_0 = 1$ and $0 \leq s_{n,j}$ for all n and j, there is a probability μ on $[0, 1]$ whose moment sequence is $\{s_n\}$. For the rest, the "if" and "only if" assertions can be proved together: the issue is to determine from the moments whether μ is absolutely continuous with respect to ν, and $d\mu/d\nu \in L_p(\nu)$. We begin by constructing an exchangeable sequence $X_1, X_2, \ldots$ of 0–1 valued random variables that satisfy (6): write P_μ for P. Define P_ν in the analogous way. Let $S_n = X_1 + \cdots + X_n$. Let $\mathcal{F}_n$ be the field generated by $X_1, \ldots, X_n$, and $\mathcal{F}$ the σ-field generated by all the X's, so $\mathcal{F}_n \uparrow \mathcal{F}$. Let H_n be the random variable whose value is $P_\mu\{S_n = j\}/P_\nu\{S_n = j\}$ on the set $\{S_n = j\}$. Then H_n is the Radon-Nikodym derivative of P_μ with respect to P_ν, both restricted to $\mathcal{F}_n$. Thus, H_n is a martingale relative to P_ν, and c_n is the p-norm of H_n relative to P_ν. By Jensen's inequality,

$$c_n \text{ in (5) are non-decreasing.} \tag{8}$$

From this point on, we use the standard martingale theory for differentiating measures. The key martingale fact is Theorem 4.1 on pp. 319–20 in Doob (1953); the application to differentiating measures is summarized in Freedman (1983, pp. 345–6): for more discussion, see Hewitt and Stromberg (1969, pp. 369–75). We conclude that

$$H_n \to H_\infty \text{ a.e. } [P_\mu + P_\nu], \tag{9}$$

with

$$H_\infty = dP_\mu/dP_\nu \tag{10}$$

for the full σ-field $\mathcal{F}$: the limit is infinite on the part of the space where P_μ is singular with respect to P_ν. Moreover,

$$c_n = [\mathrm{E}_\nu(H_n^p)]^{1/p} \uparrow [\mathrm{E}_\nu(H_\infty^p)]^{1/p}, \tag{11}$$

where E_ν denotes expectation relative to P_ν. In particular, if $\sup_n c_n \leq c < \infty$, then $H_\infty \in L_p(\nu)$ and $\|H_\infty\|_p \leq c$. On the other hand, if c_n is unbounded, then $H_\infty \notin L_p(\nu)$. The next (and last) step in the proof is perhaps worth isolating as a proposition, which writes H for H_∞.

Proposition 1. *Let $L = \lim_n S_n/n$, which exists a.e. relative to $P_\mu + P_\nu$. Let $h = d\mu/d\nu$, and $H = dP_\mu/dP_\nu$, with the understanding that $h = \infty$ on the part of the unit interval where μ is singular with respect to ν; similarly for H on its domain. Then*

(i) $P_\mu L^{-1} = \mu$.
(ii) $P_\nu L^{-1} = \nu$.
(iii) $H = h(L)$ *a.e. relative to* $P_\mu + P_\nu$.

Proof. Only claim (iii) is argued. To begin with, we impose the side condition that $\mu << \nu$. Let P_θ be the distribution when a θ-coin is tossed, so

$$P_\theta\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i},$$

the e_i being 0 or 1. Furthermore,

$$P_\mu = \int_0^1 P_\theta \, \mu(d\theta), \quad P_\nu = \int_0^1 P_\theta \, \nu(d\theta).$$

For any $A \in \mathcal{F}$,

$$\begin{aligned}
\int_A h(L)\, dP_\nu &= \int_0^1 \left(\int_A h(L)\, dP_\theta \right) \nu(d\theta) \\
&= \int_0^1 \left(\int_A h(\theta)\, dP_\theta \right) \nu(d\theta) \\
&= \int_0^1 P_\theta(A)\, h(\theta)\, \nu(d\theta) \\
&= \int_0^1 P_\theta(A)\, \mu(d\theta) = P_\mu(A).
\end{aligned}$$

The second equality holds because $P_\theta(L = \theta) = 1$ by the strong law of large numbers. The fourth equality holds by the side condition $\mu << \nu$, with $h = d\mu/d\nu$. Thus, $h(L)$ is a version of dP_μ/dP_ν. This proves claim (iii) under the side condition, but the general case follows: notice that H and h depend affinely on μ, then replace μ by $(\mu + \nu)/2$. This completes the argument, and the proof of Theorem 4. □

Remark. (i) Theorem 4 holds as stated when $p = \infty$, if we redefine c_n in (5) as

$$c_n = \max_{j=0,\ldots,n} s_{n,j}/t_{n,j}.$$

This is Corollary 3.1 in Knill (1997).

(ii) The case $p = 1$ is more problematic. We can show that $\mu << \nu$ iff the martingale H_n is uniformly P_ν-integrable, but this is little more than a restatement of the definition of absolute continuity, and uniform integrability may not be any easier to check in applications than absolute continuity.

(iii) The conditions we have considered in Theorems 2–4 are of the form

$$f_n(s_0, s_1, \ldots, s_n) \le k_n,$$

where f_n is a specified continuous function on R^{n+1}, k_n is a constant, and $s_0, s_1, \ldots$ a sequence that may—or may not—be the moment sequence of a probability that is being characterized in some way. No condition of this form can describe the moment sequences of absolute continuous probabilities, because the set of absolutely continuous probabilities is not weak-star closed.

(iv) Theorems 2–4 can be extended in a straightforward way from the unit interval to the unit cube in R^d.

(v) Hausdorff was working with finite signed measures. Theorems 2–5 can be extended to cover that case, although the interpretation of de Finetti's theorem for signed priors remains a little mysterious, at least for elderly statisticians; also see Feynman (1987). For multi-dimensional signed measures, see Knill (1997); for an application to de Finetti's theorem, see Jaynes (1986).

Local theorems

Theorem 2 can be modified if we desire only that μ should be absolutely continuous on the interval $[a, b]$, with $0 \leq a < b \leq 1$, and $d\mu/dx \leq c$ on $[a, b]$; off this interval, μ has no special features. We begin with the sufficiency part of Theorem 2, only sketching the development.

Theorem 6. *Given real numbers a, b, c with $0 \leq a < b \leq 1$ and $c > 0$, and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence by equation* (1). *There exists a probability measure μ on* $[0, 1]$ *such that*

(i) $\{s_n\}$ *is the moment sequence of μ, and*
(ii) μ *is absolutely continuous on the interval* $[a, b]$, *and*
(iii) $d\mu/dx$ *is almost everywhere bounded above by c on the interval* $[a, b]$,

if $s_0 = 1$, and $0 \leq s_{n,j}$ for all n and j, and $s_{n,j} \leq c/(n+1)$ for all n and j with $a \leq j/n \leq b$. Then μ is unique.

Here is a generalization of the sufficiency part of Theorem 4.

Theorem 7. *Given a positive real number c, and a, b with $0 \leq a < b \leq 1$, and a probability ν on* $[0, 1]$ *that assigns positive mass to* (a, b), *and a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequences $s_{n,j}$ and $t_{n,j}$ by applying equation* (1) *to μ and ν respectively. Define c_n as follows:*

$$c_n = \Big\{ \sum_{an \leq j \leq bn} t_{n,j} \Big(\frac{s_{n,j}}{t_{n,j}} \Big)^p \Big\}^{1/p}. \tag{12}$$

There exists a probability measure μ on $[0, 1]$ *such that*

(i) $\{s_n\}$ *is the moment sequence of μ, and*
(ii) μ *is absolutely continuous with respect to ν on the interval* $[a, b]$, *and*
(iii) $d\mu/dx \in L_p(\nu)$ *on the interval* $[a, b]$, *with norm at most c,*

if $s_0 = 1$, and $0 \leq s_{n,j}$ for all n and j, and $c_n \leq c$. Then μ is unique.

Proofs are straightforward, using Hausdorff's theorem to get μ and techniques described earlier in the paper to characterize $d\mu/dx$. For example, take Theorem 6. We can prove (3) for all continuous functions on the interval $[a, b]$, then for all Borel functions g on $[a, b]$ with $0 \leq g \leq 1$. The balance of the argument is unchanged. The conditions, however, are not necessary, as will be shown by example.

Example 3. To see why the upper bound in Theorem 6 cannot be a necessary condition, take $a = 0$ and $b = 1/2$. Let μ assign mass 1/2 to $[0, 1/2]$, with density bounded above by c; let μ assign the remaining mass 1/2 to $1/2 + h$. Choose n large

and even, then $h > 0$ small. Consider $P_\mu\{S_n = n/2\}$. The part of μ on $[0, 1/2]$ contributes at most $c/(n+1)$ to $P_\mu\{S_n = n/2\}$. But—if $h = 0$—the other piece of $P_\mu\{S_n = n/2\}$ is of order $1/\sqrt{n}$. If $h > 0$ is small, this other piece can therefore be much larger than $c/(n+1)$.

For Theorem 6, the necessary and sufficient upper bound condition on $s_{n,j}$ would be $s_{n,j} \le c/(n+1) + \exp(-2\delta^2 n)$ for all δ with $0 < \delta < (b-a)/2$ and all n, j with $a + \delta \le j/n \le b - \delta$. See (3.5) in Diaconis and Freedman (1990). Example 3 indicates why the term $\exp(-2\delta^2 n)$ is needed, and the restriction to $a + \delta \le j/n \le b - \delta$. The characterization of L_p densities relative to Lebesgue measure is also relatively straightforward. For other base measures, we do not have clean results.

Applications to Bayesian statistics

Theorems on moment sequences can be translated in a straightforward way into theorems characterizing the mixing measure μ in Theorem 5. We give two examples. Recall that P_θ is the distribution when a θ-coin is tossed, so

$$P_\theta\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i},$$

the e_i being 0 or 1. Furthermore,

$$P_\mu = \int_0^1 P_\theta \, \mu(d\theta). \tag{13}$$

Theorem 8. *Let X_i be 0–1 valued random variables on the probability triple $(\Omega, \mathcal{F}, P)$. Let c be a positive real number. Then $\{X_i\}$ admits the representation*

$$P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \int_0^1 \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i} f(\theta)\, d\theta$$

for all n and $e_i = 0$ or 1, and $0 \le f \le c$ a.e., iff

(i) the X_i are exchangeable, and
(ii) $P_\mu\{S_n = j\} \le c P_\lambda\{S_n = j\}$ for all $n = 0, 1, \ldots$ and $j = 0, 1, \ldots, n$, where λ is Lebesgue measure on $[0, 1]$, and $S_n = X_1 + \cdots + X_n$.

Then f is unique.

This is immediate from (7) and Theorem 2. The analog of Theorem 4 is as follows.

Theorem 9. *Let X_i be 0–1-valued random variables on the probability triple $(\Omega, \mathcal{F}, P)$. Let ν be a probability on $[0, 1]$, assigning positive mass to $(0, 1)$. Let $p > 1$ and $0 < c < \infty$. Then $\{X_i\}$ admits the representation*

$$P\{X_i = e_i \text{ for } i = 1, \ldots, n\} = \int_0^1 \theta^{\Sigma e_i}(1-\theta)^{n-\Sigma e_i} f(\theta)\, \nu(d\theta)$$

for all n and $e_i = 0$ or 1, and $f \in L_p(\nu)$ has norm at most c, iff

(i) *the X_i are exchangeable, and*
(ii) *$c_n \le c$ for all $n = 0, 1, \ldots,$ where*

$$c_n = \left[\sum_{j=0}^{n} P_\nu\{S_n = j\} \left(\frac{P_\mu\{S_n = j\}}{P_\nu\{S_n = j\}} \right)^p \right]^{1/p} \tag{14}$$

and $S_n = X_1 + \cdots + X_n$.
Then f is unique.

Theorem 9 can be extended to the case $p = \infty$ by redefining c_n as follows:

$$c_n = \max_{j=0,\ldots,n} P_\mu\{S_n = j\}/P_\nu\{S_n = j\}.$$

There are yet more general theorems characterizing partially exchangeable processes with L_p densities, in the setting of Diaconis and Freedman (1984): we will explore such results in Part II of this paper. In the abstract setting, the proofs are more transparent (although the setting itself may seem a little strange).

Monotone densities

In some applications, it is desired to characterize monotone densities in terms of their moments; see, for instance, Diaconis and Kemperman (1996). Theorem 10 gives a result for densities that are non-decreasing. We will need the following lemma, which expresses a monotone function as a mixture of the extreme step functions.

Lemma 2. *Let F be a non-negative, right-continuous, non-decreasing function on $[0, 1)$; we allow $F(0) > 0$ and $F(1-) = \infty$. Let $f_\theta(x) = 0$ for $0 \le x < \theta$ and $f_\theta(x) = 1/(1-\theta)$ for $\theta \le x < 1$, so f_θ is a probability density for $0 \le \theta < 1$. Then*

$$F = \int_{[0,1)} f_\theta \, \nu(d\theta),$$

where the measure ν on $[0, 1)$ is defined as follows: $\nu(d\theta) = (1-\theta)F(d\theta)$, with $F(d\theta)$ assigning mass $F(0)$ to 0. Finally, the total mass in ν is $\int_0^1 F(x)\,dx$.

Proof. The calculation will seem trite, but it is easy to get lost if you start at the wrong place. Let $H_\theta = 0$ on $[0, \theta)$ and $H_\theta = 1$ on $[\theta, 1)$. Then

$$F(x) = \int_{[0,1)} H_\theta(x)\, F(d\theta) = \int_{[0,1)} f_\theta(x)\, (1-\theta) F(d\theta) = \int_{[0,1)} f_\theta(x)\, \nu(d\theta).$$

To evaluate the mass in ν, integrate over $x \in [0, 1)$. The proof is complete. □

Theorem 10. *Given a sequence $s_0, s_1, \ldots$ of real numbers, define the auxiliary sequence $s_{n,j}$ by equation* (1). *There exists a probability measure μ on $[0, 1]$ such that*

(i) $\{s_n\}$ *is the moment sequence of* μ, *and*
(ii) μ *is absolutely continuous on* $[0, 1)$, *and*
(iii) $d\mu/dx$ *is non-decreasing on* $[0, 1)$,

if and only if $s_0 = 1$, *and* $0 \le s_{n,j}$ *for all* n *and* j, *and* $s_{n,j}$ *is nondecreasing in* j *for all* n. *The probability* μ *has a possible atom at* 1, *but* $\mu\{1\} = 0$ *iff* $s_n \to 0$.

Proof. Suppose μ satisfies conditions (i), (ii), and (iii). Then μ is a convex combination of point mass at 1, and an absolutely continuous probability on $[0, 1]$ with a non-decreasing density. If $\mu\{1\} = 1$, it is clear that $s_{n,j}$ is non-decreasing with j. Suppose on the other hand that μ is absolutely continuous on $[0, 1]$ and $d\mu/dx$ is non-decreasing. As in Lemma 2,

$$d\mu/dx = \int_{[0,1)} f_\theta \, \nu(d\theta).$$

(In this application, ν is a probability measure.)

Since $s_{n,j}$ is affine in μ by Lemma 1a, it suffices to consider the θ's one at a time, i.e., we can take ν to be point mass at θ. Let $0 \le j < n$. We claim that $s_{n,j} \le s_{n,j+1}$, that is,

$$\binom{n}{j} \int_0^1 x^j (1-x)^{n-j} f_\theta(x)\, dx \le \binom{n}{j+1} \int_0^1 x^{j+1} (1-x)^{n-j-1} f_\theta(x)\, dx, \tag{15}$$

which is to say,

$$(j+1) \int_\theta^1 x^j (1-x)^{n-j}\, dx \le (n-j) \int_\theta^1 x^{j+1} (1-x)^{n-j-1}\, dx. \tag{16}$$

Let $G(\theta)$ be the right hand side of (16) minus the left hand side, namely,

$$G(\theta) = \int_\theta^1 x^j (1-x)^{n-j-1} g(x)\, dx,$$

where

$$g(x) = (n-j)x - (j+1)(1-x).$$

Now

$$G'(\theta) = \theta^j (1-\theta)^{n-j-1} h(\theta),$$

where

$$h(\theta) = -g(\theta) = (j+1)(1-\theta) - (n-j)\theta.$$

Clearly, $h(\theta) > 0$ for $0 \le \theta < (j+1)/(n+1)$ and $h(\theta) < 0$ for $(j+1)/(n+1) < \theta \le 1$. Thus, G increases from 0 at 0—see Lemma 1c—to its maximum at $(j+1)/(n+1)$, and then decreases to 0 at 1. In short, $G > 0$ except at 0 and 1, where G vanishes. Thus, (16) holds for $0 \le \theta \le 1$, and (15) must hold for $0 \le \theta < 1$, completing the proof of the "only if" part of the theorem. The converse follows from Proposition 2 below, with $p_{n,j} = s_{n,j}$. The convergence of μ_n is discussed in the remarks following the proposition. □

Proposition 2. *Let the probability μ_n on $[0, 1]$ assign mass $p_{n,j}$ to j/n for $j = 0, 1, \ldots, n$, with $0 \le p_{n,0} \le p_{n,1} \le \cdots \le p_{n,n}$ and $\sum_{j=0}^{n} p_{n,j} = 1$. Suppose $\mu_n \to \mu$ weak-star. Let F be the distribution function of μ. Then F is convex on $[0, 1]$, hence absolutely continuous on $[0, 1)$ with nondecreasing density F'. There is a possible atom at 1.*

Proof. Take the convolution of μ_n with the uniform distribution on $[-\frac{1}{2n}, \frac{1}{2n}]$, in effect replacing the point masses with their histogram. The resulting measure has distribution function F_n which is convex—because F_n' is monotone—and still converges weak-star to F. Let D be the set of discontinuity points of F. Then $D \cup D/2 \cup D/3 \cup \cdots$ is countable. So, there are small positive h with $jh \in D$ for no integer j: after all, $jh \in D$ iff $h \in D/j$. Next, F_n converges pointwise to F on the h-skeleton $h, 2h, \ldots$, because F is continuous there. Since F_n is convex on this skeleton, so is F. But h can be arbitrarily small. Therefore, F is convex on $(0, 1)$. In particular, F is continuous on $(0, 1)$, even absolutely continuous, and its density F' is increasing. Suppose by way of contradiction that 0 were an atom with mass $\delta > 0$. For any $x, h > 0$ with $0 < x < 1 - h$, we would have $\mu[x, x + h] = \lim_n F_n(x + h) - F_n(x) \ge \limsup_n F_n(h) - F_n(0) \ge \delta$, which is impossible; the first inequality holds because F_n' is monotone; the second, because $p_{0,n} \le 1/(n + 1)$ so $F_n(0) - F_n(-h) \to 0$ while $\mu\{0\} = \delta$. Thus, F is continuous even at 0, with $F(0) = 0$. This finishes the proof of Proposition 2, and hence of Theorem 10. □

Remarks.

(i) Decreasing densities can be characterized in a similar way, although the possible atom moves to 0, and can be excluded by requiring $s_{n,0} \to 0$.

(ii) The existence of the density in Theorem 10 follows from the monotonicity of $s_{n,j}$, but the density need not be bounded.

(iii) Why does μ_n converge? Hausdorff proved Theorem 1 by showing directly that μ_n converges weak-star to the desired μ: see Feller (1971, pp. 225–26). For us, it may seem more natural to prove the relevant law of large numbers. The convergence of μ_n would follow, along with Hausdorff's moment theorem, the convergence of the Bernstein polynomials, and de Finetti's theorem. In essence, that is the path followed by de Finetti (1937). Compactness arguments are also feasible.

(iv) Theorem 10 completes Bayes' observation that a uniform density corresponds to a uniform distribution for S_n: the uniform density is non-decreasing and non-increasing, so the resulting distribution of S_n has the same features. Of course, there are familiar arguments that are more direct: see Lemma 1 and the remarks that follow it.

(v) Suppose μ is absolutely continuous on $[0, 1)$, and $d\mu/dx$ is non-decreasing on $[0, 1)$. Unless $d\mu/dx$ is constant, $s_{n,j}$ will be strictly increasing with j. Indeed, the inequality in (16) is strict unless $\theta = 0$ or 1; the inequality in (15) is therefore strict unless $\theta = 0$, corresponding to a density that is constant. On the other hand, if μ has an atom at 1, then $s_{n,n-1} < s_{n,n}$.

Historical notes

Hausdorff

Hausdorff's work on the moment problem was motivated by summability theory (Hausdorff, 1921, 1923). In brief, let $S = \{s_{n,j} : n = 0, 1, \ldots, j = 0, \ldots, n\}$ be a triangular matrix of real numbers. The "S-limit" of a sequence $\{x_i\}$ is $\lim_n \sum_{j=0}^n s_{n,j} x_j$. A summability method S is "regular" if $\lim x_i = x_\infty$ implies that the S-limit is x_∞. Familiar examples include Cesàro's method, where $s_{n,j} = 1/(n+1)$, and Euler's E_p method with

$$s_{n,j} = \binom{n}{j} p^j (1-p)^{n-j}.$$

Hausdorff introduced a more general scheme, defining

$$s_{n,j} = \binom{n}{j} \int_0^1 p^j (1-p)^{n-j} \, \mu(dp) \tag{17}$$

where μ is a finite signed measure on $[0, 1]$. For instance, setting μ to Lebesgue measure gives us Cesàro's method: see Lemma 1b. If μ is point mass at p, we get E_p. Among many other things, Hausdorff showed that a summability method defined by (17) is regular iff $\mu\{0\} = 0$ and $\mu(0, 1] = 1$; this is more or less obvious from (7). However, μ need not be a probability measure: its negative part need not vanish. Methods defined by (17) are now called "Hausdorff methods." For additional discussion, see Widder (1946) or Hardy (1949).

Some notes on Hausdorff (1923) may be of interest. The auxiliary sequence, with the binomial coefficients, is introduced in equation (5) on p. 223; the positivity condition is (A) on the same page. The solution to the moment problem is Satz I on p. 226. The condition for an L_p density is (C) on p. 234, and the theorem is Satz III on p. 236. The condition for L_∞ is (D) on the same page, and the solution to the Markov moment problem is Satz IV on p. 237. The hitherto-unmentioned Satz II on p. 232 characterizes moment sequences of finite signed measures: his necessary and sufficient condition (B) is, in our notation, $\sup_n \sum_j |s_{n,j}| < \infty$.

The Russian School

Solutions to the Markov moment problem, and similar results for the half-line and the whole line, were among the great achievements of the Russian school. Perhaps the history begins with Chebychev, who gave a rigorous proof of the Central Limit Theorem using the method of moments, with connections to the theory of continued fractions, orthogonal polynomials, and numerical quadrature. His student Markov formulated the moment problem we have been discussing (along with many other contributions in other areas).

Let $\{s_n\}$ be a given sequence of real numbers, and c a given positive real. When is $s_n = \int_0^1 x^n f(x)\, dx$ for all n, with f a probability density bounded above by c?

To answer this question, Markov expanded

$$\exp\Big[\frac{1}{c}\Big(\frac{s_0}{z}+\frac{s_1}{z^2}+\frac{s_2}{z^3}+\frac{s_3}{z^4}+\cdots\Big)\Big] \tag{18}$$

as a continued fraction, and showed that positivity of certain coefficients was a necessary condition. The condition turned out to be sufficient as well. There were later developments by Ahiezer and Krein (1962), and Krein and Nudelman (1977). One theorem in Ahiezer and Krein (1962, p. 71) can be stated this way: $s_n = \int_0^1 x^n f(x)\,dx$ for all n, with $0 \le f \le c$ a.e., iff t_n satisfies Hausdorff's condition, where t_n is defined by a formal series expansion of (18) in powers of $1/z$:

$$\exp\Big[\frac{1}{c}\Big(\frac{s_0}{z}+\frac{s_1}{z^2}+\frac{s_2}{z^3}+\frac{s_3}{z^4}+\cdots\Big)\Big] = 1+\frac{t_1}{z}+\frac{t_2}{z^2}+\frac{t_3}{z^3}+\cdots \tag{19}$$

Of course, the t_n are polynomial functions of s_n/c. For example,

$$t_1 = \frac{s_0}{c}, \quad t_2 = \frac{s_1}{c}+\frac{1}{2}\frac{s_0^2}{c^2}, \quad t_3 = \frac{s_2}{c}+\frac{s_0 s_1}{c^2}+\frac{1}{6}\frac{s_0^3}{c^3}.$$

In general,

$$t_n = \frac{1}{n!}\sum_{\pi}\prod_{j=1}^{n}(j s_{j-1}/c)^{a_j(\pi)} \tag{20}$$

where π runs through the permutations of length n, and $a_j(\pi)$ is the number of cycles in π of length j. Here, $s_0 = \int f$: if $s_0 = 1$, then f is a probability density.

Sergei Kerov made several remarkable contributions to this theory. For instance, (19) sets up a one-to-one correspondence between the moments $\{s_n\}$ of a density bounded by c, and the moments $\{t_n\}$ of an auxiliary measure ν on $[0, 1]$. Given f, Kerov showed how to pick a random point from ν, by generating a nested sequence of random intervals

$$[0,1] \supset [X_1, Y_1] \supset [X_2, Y_2] \supset \cdots$$

that shrink to a point. Despite the complexity of (19), Kerov's algorithm is elegance itself. At stage $n+1$, pick a point U at random in $[X_n, Y_n]$. Then flip a coin that lands heads with probability $f(U)/c$, or tails with the remaining probability $1 - [f(U)/c]$. If the coin lands heads, $X_{n+1} = U$ and $Y_{n+1} = Y_n$. But if the coin lands tails, $X_{n+1} = X_n$ and $Y_{n+1} = U$. Probabilities have to be bounded between 0 and 1: that is where the condition $0 \le f \le c$ comes in.

Kerov found striking connections between his algorithm and Young tableaux, as well as eigenvalues of random matrices, and the zeroes of orthogonal polynomials. Recently, expansions connected to the Markov moment problem—like (19) and (20)—have found applications in Bayesian non-parametric statistics: Cifarelli and Regazzini (1990), Diaconis and Kemperman (1996).

Acknowledgements. We thank Christian Berg for suggesting a beautiful alternative proof for the "if" part of Theorem 2. The s_n are moments of a probability on [0, 1], call it μ, by Hausdorff's theorem. Let λ be Lebesgue measure on [0, 1], and $\{t_n\}$ the moment sequence of $c\lambda - \mu$, with auxiliary sequence $\{t_{n,j}\}$ defined by the analog of (1). Then

$$t_{n,j} = \binom{n}{j} \int_0^1 x^j (1-x)^{n-j} d(c\lambda - \mu) = c/(n+1) - s_{n,j}$$

by Lemma 1, so $\{t_n\}$ satisfies the Hausdorff condition,

$$(-1)^{n-j} \binom{n}{j} \Delta^{n-j} t_j \geq 0,$$

although $t_0 = 1$ is unlikely. Hence, $c\lambda \geq \mu$, again by Hausdorff's theorem. We also thank Jon McAuliffe for a number of helpful comments, and a very careful anonymous referee.

References

Ahiezer, N.I., Krein, M.: Some Questions in the Theory of Moments. Amer. Math. Soc. Providence, 1962

Bayes, T.: An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London **53**, 370–418 (1764)

Cifarelli, D., Regazzini, E.: Distribution functions of means of a Dirichlet process. Ann. Statist. **18**, 429–442 (1990)

de Finetti, B.: Funzione caratteristica di un fenomeno aleatorio. Atti della R. Academia Nationale dei Lineii Ser. 6, Memorie, classe di Scienze, Fisiche, Matematiche e Naturali **4**, 251–99 (1931)

de Finetti, B.: La prévision: ses lois logiques ses sources subjectives. Ann. Inst H. Poincaré **7**, 1–68 (1937); Translated in H. Kyburg, H. Smokler, (eds.), Studies in Subjective Probability. Wiley, New York, 1964

Diaconis, P., Freedman, D.: Partial exchangeability and sufficiency. In J. K. Ghosh, J. Roy, (eds.), Proc. Indian Statist. Assoc. Golden Jubilee: Applications And New Directions. Indian Statist. Assoc., Calcutta, 1984, pp. 205–236

Diaconis, P., Freedman, D.: On the uniform consistency of Bayes estimates for multinomial probabilities. Ann. Statist. **18**, 1317–27 (1990)

Diaconis, P., Kemperman, J.: Some new tools for Dirichlet priors. In J. Bernardo et al, (eds.), Bayesian Statistics 5, Oxford University Press, 1996, pp. 97–106

Doob, J.L.: Stochastic Processes. Wiley, New York. Reprinted 1990, Wiley Classics Library, 1953

Feller, W.: An Introduction to Probability Theory and its Applications. Vol. II, Second edition, Wiley, New York, 1971

Feynman, R.: Negative probabilities. In: B. J. Hiley, F. D. Peat, (eds.), Quantum Implications. Routledge & Kegan-Paul, 1987

Freedman, D.A.: Markov Chains. Springer, New York. First published in 1971 by Holden Day, San Francisco, 1983

Grabner, P.J., Prodinger, H.: Asymptotic analysis of the moments of the Cantor distribution. Statistics & Probability Letters **26**, 243–8 (1996)

Hardy, G.H.: Divergent Series. Oxford. Reprinted 1991 by Chelsea, New York, 1949

Hausdorff, F.: Summationsmethoden und Momentfolgen I, II. Math. Zeitschrift **9**, 74–109, 280–290 (1921)

Hausdorff, F.: Momentprobleme für ein endliches intervall. Math. Zeitschrift **16**, 220–48 (1923)
Hewitt, E., Savage, L.J.: Symmetric measures on cartesian products. TAMS **80**, 470–501 (1955)
Hewitt, E., Stromberg, K.: Real and Abstract Analysis. Springer, New York, 1969
Jaynes, E.I.: Some applications and extensions of the de Finetti representation theorem. In: Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. P. K. Goel and A. Zellner, (eds.), North-Holland, Amsterdam, 1986, pp. 31–42
Kerov, S.: Transition probabilities for continual Young diagrams and the Markov moment problem. Functional Analysis and its Applications **27**, 104–117 (1993)
Kerov, S.: Interlacing Measures. Amer. Math. Soc. Transl. Ser. 2 **181**, 35–83 (1998); Advances in the Mathematical Sciences, 35, G. I. Olshanski, (ed.)
Knill, O.: On Hausdorff's moment problem in higher dimensions. Technical report, Department of Mathematics, Harvard University, 1997 http://abel.math.harvard.edu/~knill/preprints/
Krein, M.G., Nudelman, A.: The Markov Moment Problem and Extremal Problems. Amer. Math. Soc. Providence, 1977
Lorentz, C.: Bernstein Polynomials. University of Toronto Press, Toronto, 1966
Shohat, J., Tamarkin, J.: The Problem of Moments. Amer. Math. Soc. Providence, 1943
Stigler, S.M.: The History of Statistics. Harvard University Press, 1986
Widder, D.V.: The Laplace Transform. First printing, 1941, Princeton University Press, 1946