Moment information and entropy evaluation for probability densities

Mariyan Mileva a, Pierluigi Novi Inverardi b, Aldo Tagliani b,*

a Dept. of Informatics and Statistics, University of Food Technologies, 4002 Plovdiv, Bulgaria
b Dept. of Computer and Management Sciences, University of Trento, 38100 Trento, Italy

A R T I C L E   I N F O

Keywords:
Entropy
Entropy convergence
Hausdorff moment problem
Kullback–Leibler distance
Maximum entropy

A B S T R A C T

How much information does the sequence of integer moments carry about the corresponding unknown absolutely continuous distribution? We prove that a reliable evaluation of the corresponding Shannon entropy can be done by exploiting some known theoretical results on the entropy convergence, uniquely involving exact moments without solving the underlying moment problem. All the procedure essentially rests on the solution of linear systems, with nearly singular matrices, and hence it requires both calculations in high precision and a pre-conditioning technique. Numerical examples are provided to support the theoretical results.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Let be X an absolutely continuous random variable with support D. The moment problem consists of recovering an unknown probability density function (pdf) \( f(x) \in L^2_0 \) from the knowledge of its associated sequence \( \{\mu_j\}_{j=0}^M \) of integer moments, with \( M \) finite or infinite, where \( \mu_j = \mathbb{E}(X^j) = \int_D x^j f(x) \, dx, \quad j \geq 0, \quad \mu_0 = 1. \) In actual practice, the problem consists of determining approximations to the unknown density function \( f(x) \) from a finite collection \( \{\mu_j\}_{j=0}^M \) of its integer moments. In such a case the moment problem is called “truncated moment problem”, else, if \( M \) is infinite the problem is called “full moment problem”. If \( D \) is bounded, typically \( D = [0,1] \), we say “Hausdorff moment problem”, whilst if \( D = \mathbb{R}^+ \) Stieltjes moment problem and if \( D = \mathbb{R} \) Hamburger moment problem. In Hausdorff case the infinite sequence of moments determines a unique distribution (we say determinate moment problem), unlike Stieltjes and Hamburger cases where supplementary conditions are requested to guarantee an unique distribution. The above terminology and general results about existence and determinacy are found in the classical book of Shohat and Tamarkin [1]. Once assigned the first moments \( \{\mu_j\}_{j=0}^M \) we define the \( M \)th moment space \( \mathcal{V}_M \subset \mathbb{R}^M \) as the convex hull of the curve \( \{x, x^2, \ldots, x^M\}, \quad x \in D; \) in Hausdorff case \( \mathcal{V}_M \) is closed, bounded and convex [2, Theorem 7.2] and its geometrical properties are well described in [2].

In the last decade several papers have been devoted on determining how the partial information contained in the finite sequence of integer moments can be used to describe various distributional characteristics and properties. Lindsay and Basak [3] show tail bounds for an unknown distribution in terms of moments by exploiting a classical bounding result due to Akhiezer [4] and Shohat and Tamarkin [1]. Goria and Tagliani [5] present a moment bound for the absolute difference between two distributions, based on maximum entropy (MaxEnt) method; Gavrilidis and Athanassoulis [6] and Gavrilidis [7] obtain some results about the separation of the main mass interval, the tail interval and the position of the mode. In some recent papers Mnatsakanov [8,9] provides a procedure to recover a probability density function (pdf) \( f(x) \) and the associated distribution function \( F(x) \) directly from the associated moments without resorting to an approximation \( \hat{f}(x) \) or \( f_x(x) \) or...
\(F^{(\text{app})}(x)\) of \(f(x)\); in the framework of the Hausdorff moment problem Mntaksanov aims to construct a stable approximation \(f^{(\text{app})}(x)\) of \(f(x)\) from its moments sequence and proves some asymptotic properties of the approximant and finally derives the uniform and \(L_1\)-rates of convergence.

All the previous results come from an obvious fact that is, in a determinate moment problem the information content about a distribution is spread on the infinite sequence of its moments. In the present paper we run along the guidelines of the above quoted papers. In more details, we tackle the issue to determining the Shannon-entropy of a distribution with support \(D = [0,1]\) and admitting pdf \(f(x)\), without solving the inverse moment problem, avoiding consequently all the instability drawbacks to the solution of inverse problem. A viable route to calculating Shannon-entropy

\[
H[f] = -\int_0^1 f(x) \ln f(x) dx
\]

consists in approximating \(f(x)\) by a density \(f^{(\text{app})}(x)\), constrained by a finite set of moments. Then

\[
\int_0^1 x^j f^{(\text{app})}(x) dx = \int_0^1 x^j f(x) dx, \quad j = 0, \ldots, M \\
\text{and } H[f] \text{ is replaced by } H[f^{(\text{app})}] = -\int_0^1 f^{(\text{app})}(x) \ln f^{(\text{app})}(x) dx.
\]

The procedure that we propose here requires an accurate recovering of \(f(x)\) through \(f^{(\text{app})}(x)\) by involving a high number \(N\) of moments \(\{\mu_j\}_{j=0}^M\). In general, high \(N\) values arise ill-conditioning issues in the involved Hankel matrices. In Hausdorff case, for instance, \([10]\) proved that all Hankel matrices generated by moments of positive density functions are conditioned essentially the same as Hilbert matrices. Then each order \(N\) Hankel matrix has condition number which grows at exponentially rate \(\approx e^{3.525M}\) for large \(M\). Condition number of Hankel matrices generated by moments arising from a distribution with support \(\mathbb{R}^+\) or \(\mathbb{R}\) may be found in \([11,12]\). All the latter matrices have condition number growing at exponentially rate. Then, to take ill conditioning into account, only few moments have to be involved with a consequent poor estimate of \(R\), unless a suitable preconditioner is available, as in Hausdorff case \([10]\). Indeed, although the infinite sequence of moments \(\{\mu_j\}_{j=0}^\infty\) completely determines the unknown distribution, the information content carried out by each moment \(\mu_j\) depends on the sequence itself; in general, it could be poor or even insignificant when the corresponding distribution exhibits fat tails as well as few moments may sometimes characterize completely the distribution: in such a case they coincide with the “characterizing moments”. We recall that the characterizing moments of a probability distribution are the moment constraints subject to which the probability distribution is obtained as the MaxEnt probability distribution \([13, p. 135]\). They are able to synthesize all the relevant information on the distribution and characterize it.

For instance, if \(D = [0,1]\), when \(M\) takes a high value from Taylor expansion one has \(x^M \approx \sum_{n \in \mathbb{N}} (x - \frac{1}{2})^n\). Taking the expected value one gets

\[
\mu_M = E(X^M) \approx \sum_{n = 0}^{\infty} E(X - \frac{1}{2})^n = \sum_{n = 0}^{\infty} \sum_{j = 0}^{n} \binom{n}{j} \left(-\frac{1}{2}\right)^{n-j} \mu_j.
\]

Equivalently, higher moments may be obtained combining lower order moments. Heuristically speaking, different powers of \(x\) differ very little from each other if \(x\) is restricted by \(0 < x < 1\) and the exponents are large. Then a high number of moments has to be involved in an accurate recovering with the consequent risk to run into numerical instability.

An upper bound for entropy \(H[f]\) uniquely in terms of first \(M + 1\) moments \(\mu = (\mu_0, \ldots, \mu_M)\) had been provided \([14, Eq. (3.8)]\). If the moment vector \(\mu^j = \{\mu_j\}_{j=0}^M\) of the uniform distribution in \([0,1]\) is considered, it holds

\[
H[f] \leq \inf_{\mu^j} \left[ -\frac{\|\mu - \mu^j\|_2^2}{2\|\max(\mu, \mu^j)\|_1} \right].
\]

The bound (\ref{eq:bound}) is tight whenever the moment vector \(\mu\) is close enough to the moment vector \(\mu^j\), or equivalently \(H[f] \approx 0\) (see \([14]\) for details).

Exploiting previous results of convergence characterizing MaxEnt approximations of \(f(x)\) we try to bypass that instability and we illustrate a viable way to calculate the entropy of the underlying distribution uniquely in terms of moments. The needed procedure requires uniquely the solution of linear systems and the evaluation of definite integrals (\ref{eq:entropy}) and (\ref{eq:maxent}) below. An accurate evaluation of \(H[f]\) requires both a high number of moments and a high precision computational procedure jointly with a suitable preconditioner for the matrix involved in the linear system.

The paper is organized as follows. In Section 2 the MaxEnt problem is formulated and convergence theorems, as entropy, directed divergence and almost everywhere are reviewed. Based upon above theorems an approximate procedure for entropy evaluation is formulated. In Section 3 a suitable preconditioner for solving ill conditioned linear systems is illustrated, as well as numerical examples validating the above approximate method. In Section 4 expected values to be used in MaxEnt setup alternative to integer moments are suggested. An extension to Stieltjes or Hamburger moment problems of the above procedure for approximate entropy evaluation is discussed.

2. Formulation of the problem

Let us consider a random variable \(X\) with support \(D = [0,1]\), admitting pdf \(f(x)\) with moments \(\{\mu_j\}_{j=0}^\infty\) and underlying determinate moment problem (so called Hausdorff moment problem). For practical purposes, only a finite sequence \(\{\mu_j\}_{j=0}^M\) has to be considered. Then there exist an infinite set of densities \(f^{(\text{app})}(x)\) compatible with the information carried from the finite moment sequence and hence a choice criterion is required. The MaxEnt approach is widely used in the practice and provides the approximation of \(f(x)\) \([13]\).
\[ f^{\text{app}}(x) =: f_M(x) = \exp \left( - \sum_{j=0}^{M} \lambda_j x^j \right), \] (2.1)

constrained by

\[ \int_{D} x^j f_M(x) dx = \mu_j, \quad j = 0, \ldots, M. \] (2.2)

with Shannon entropy

\[ H[f_M] = - \int_{D} f_M(x) \ln f_M(x) dx = \sum_{j=0}^{M} \lambda_j \mu_j \] (2.3)

and 0 \( \geq H[f_1] \geq \cdots \geq H[f_M] \geq \cdots \geq H[f] \) as is known from the MaxEnt context. Lagrange multipliers \( (\lambda_0, \ldots, \lambda_M) \) are evaluated by solving the minimization problem [13]

\[ \{ \lambda_j \}_{j=1}^{M} : \min_{\lambda_1, \ldots, \lambda_M} \left[ \ln \left( \int_{D} \exp \left( - \sum_{j=0}^{M} \lambda_j x^j \right) dx \right) + \sum_{j=1}^{M} \lambda_j \mu_j \right] \] (2.4)

and

\[ \lambda_0 = \ln \int_{D} \exp \left( - \sum_{j=1}^{M} \lambda_j x^j \right) dx. \]

Then a direct evaluation of \( H[f_M] \) is obtained by \( H[f_M] = \sum_{j=0}^{M} \lambda_j \mu_j \). A moment problem solution produces instability issues when a high number \( M \) of moments is involved. On the other hand the entropy \( H[f_M] \) of \( f_M(x) \) represents one of the main indicators to test the reliability of the approximation. Combining entropy convergence (Theorem 1 below) and the obvious sequence of inequalities \( 0 \geq H[f_1] \geq \cdots \geq H[f_M] \geq \cdots \geq H[f] \) the criterion for the choice of the number \( M \) of moments that yield the best approximation \( f_M(x) \) before incurring in numerical instability (equivalently, the sequence \{\( H[f_j] \)\} fails to be monotonic decreasing) should be

\[ M \equiv j : \min_{j} H[f_j]. \]

This fact is due to conditioning of \( H[f_M](\mu_M) \) (which we consider depending on \( \mu_M \) only, whilst the remaining moments \( (\mu_0, \ldots, \mu_{M-1}) \) are held fixed), whose value is given by Tagliani [15]

\[ \text{Cond}(H[f_M]) = \frac{|H[f_M](\mu_M + \Delta \mu_M) - H[f_M](\mu_M)|}{H[f_M]} \geq \epsilon \frac{(\Delta \mu_M)^2}{H[f]} 2^{4(M-3)}, \]

where \( \epsilon = |\Delta \mu_M|/\mu_M \) may be identified with the machine error. Then the condition of \( H[f_M] \) grows exponentially and justifies the use of a low number \( M \) of integer moments in recovering \( f_M(x) \) by (2.4). Such a drawback may be bypassed by resorting to some known theoretical results, arising an alternative procedure to calculate the entropy.

In the context of MaxEnt the following convergence result holds [15].

**Theorem 1.** Whenever the underlying moment problem is determinate, MaxEnt approximations converge in entropy, as \( M \) increases, to \( f(x) \), i.e.

\[ \lim_{M \to \infty} H[f_M] = H[f]. \] (2.5)

Denote \( \|f, f_M\| = \int_{D} f(x) \ln \frac{f(x)}{f_M(x)} dx \) the Kullback–Leibler distance between \( f(x) \) and \( f_M(x) \); since \( f(x) \) and \( f_M(x) \) share same first \( M \) moments \( \{\mu_j\}_{j=0}^{M} \) it follows [13]

\[ H[f_M] - H[f] = \int_{D} f(x) \ln \frac{f(x)}{f_M(x)} dx. \] (2.6)

Combining (2.5) with (2.6) it follows that

\[ \lim_{M \to \infty} H[f_M] - H[f] = \lim_{M \to \infty} \int_{D} f(x) \ln \frac{f(x)}{f_M(x)} dx = 0. \] (2.7)

Invoking the following result.

**Theorem 2.** [16, p. 252]

\[ \|f, f_M\| \geq 0, \]

with equality if and only if \( f \) and \( f_M \) coincide almost everywhere.
Comparing and then being the procedure (2.9) for calculating or here the authors [17] observe what is remarkable about (2.11) is that, not only they are exact, but they exhibit a strikingly simple bilinear form, despite the strong non-linearity of the involved pdf moments has to be considered and approximations to the closure relations interesting the Fluid Dynamics, i.e. to express the higher order moments in terms of lower order ones and Lagrange multipliers (2.12) or (2.13) leads to inaccurate results, as clarified by their numerical analysis. Theorem 1 leads to inaccurate results, as clarified by their numerical analysis. Theorem 1, one has.

\[
\mu_j = \mu_j(f) = \int_0^1 x f(x) \, dx = \mu_j(f^{\text{app}}) = \int_0^1 x f^{\text{app}}(x) \, dx, \quad j \geq M. 
\]

Once calculated \( f_M(x) = \exp \left( -\sum_{k=1}^M x_k \right) \), higher moments \( \mu_j(f_M), j > M \) may be found in terms of \( \{\mu_j\}_j^{M} \) and \( \{\lambda_j\}_j^{M} \) integrating by parts (2.2) with the following result

\[
\sum_{k=1}^M k \lambda_k (\mu_k - \mu_{k-1}) = 1 - (j + 1) \mu_j, \quad j = 1, \ldots, M. 
\]

The linear system (2.9) admits solution being an identity relating \( \lambda = (\lambda_1, \ldots, \lambda_M) \) with \( \{\mu_j(f_M)\}_{j=M+1}^{2M} \). From (2.2) (with \( j = 0 \)), \( \lambda_0 \) may be obtained integrating \( f_M(x) \) to one

\[
\lambda_0 = \ln \int_0^1 \exp \left( -\sum_{j=1}^M x_j \right) \, dx. 
\]

Remark 2.1. (2.9) is not new in literature. Similar relationships had been previously obtained by several authors. We restrict ourselves to recall the system of moment equations obtained in [17,18] for the Hamburger case (here an even number \( M \) of moments has to be considered and \( D \equiv \mathbb{R} \))

\[
\text{Han}_{2M} \cdot [\lambda_1, 2\lambda_2, \ldots, M\lambda_M]^T = [0, 1, \mu_0, 2\mu_1, \ldots, (M + 1)\mu_M]^T, 
\]

with rectangular matrix \( \text{Han}_{2M} \)

\[
\text{Han}_{2M} = \begin{bmatrix} \mu_0 & \cdots & \mu_{M-1} \\ \vdots & \vdots & \vdots \\ \mu_{M+1} & \cdots & \mu_{2M} \end{bmatrix}. 
\]

Here the authors [17] observe what is remarkable about (2.11) is that, not only they are exact, but they exhibit a strikingly simple bilinear form, despite the strong non-linearity of the involved pdf. From the third equation on, (2.11) involves moments of higher order than those which were constrained. Therefore, these equations can be used to obtain successful approximations to the closure relations interesting the Fluid Dynamics, i.e. to express the higher order moments in terms of the lower order ones and Lagrange multipliers (\( \lambda_1, \ldots, \lambda_M \)).

Remark 2.2. Alternatively, \( \lambda_0 \) may be obtained directly from \( \{\mu_j\}_j^{M} \) and \( \{\lambda_j\}_j^{M} \) integrating by parts (2.2) (with \( j = 0 \))

\[
\lambda_0 = -\ln \left[ \frac{\sum_{j=1}^M j \lambda_j \mu_{j-1}}{1 - \exp \left( -\sum_{j=1}^M \lambda_j \right)} \right]. 
\]

or

\[
\lambda_0 = -\ln \left[ \frac{1 - \sum_{j=1}^M \mu_j \lambda_j}{\exp \left( \sum_{j=1}^M \lambda_j \right)} \right]. 
\]

Being the procedure (2.9) for calculating \( \lambda = (\lambda_1, \ldots, \lambda_M) \) only approximate, the use of (2.12) or (2.13) leads to inaccurate results, as clarified by their numerical analysis. The procedure (2.9) for calculating \( \lambda = (\lambda_1, \ldots, \lambda_M) \) is approximate in the following sense: for a fixed \( M, (2.9) \) involves \( \mu_j(f_M) \), which are unknown for \( j = M + 1, \ldots, 2M \), whilst \( \mu_j(f_M) = \mu_j(f), j = 1, \ldots, M \) are known. For practical purposes such \( \mu_j(f_M), j = M + 1, \ldots, 2M \) are replaced by the assigned \( \mu_j(f) \), with error decreasing as \( M \) increases, given as follows.

From Pinsker inequality [19, p. 390] and (2.6)

\[
H[f_M] - H[f] = \int_D f(x) \ln \frac{f(x)}{f_M(x)} \, dx \geq \frac{1}{2} \left( \int_0^1 [f_M(x) - f(x)]^2 \, dx \right)^2 
\]

and then

\[
\int_0^1 [f_M(x) - f(x)]^2 \, dx \leq 2(H[f_M] - H[f]). 
\]

Comparing \( \mu_j(f) \) and \( \mu_j(f_M), j = M + 1, \ldots, 2M \), calling \( \Delta \mu_j = \mu_j(f_M) - \mu_j(f) \) and taking into account Theorem 1, one has
\[
|\Delta k| = |\mu_k(f_M) - \mu_l(f)| = \left| \int_0^1 x f_M(x) - f(x) \, dx \right| \leq \int_0^1 x f_M(x) - f(x) \, dx \leq \int_0^1 |f_M(x) - f(x)| \, dx \\
\leq \sqrt{2[H[f_M] - H[f]]} \to 0. 
\]  
(2.14)

Then \(\mu_k(f_M)\) closely approximates \(\mu(f)\), \(j = M + 1, \ldots, 2M\), as \(M\) increases. For high \(M\) values, after replacing \(\mu_k(f_M)\) with \(\mu_j(f)\), \(j = M + 1, \ldots, 2M\), the linear system (2.9) continues to admit solutions as a consequence of Theorem 2.

From a combined use of (2.9) and (2.10) we have as follows. We call.

(i) \(H[f_M] = \lambda_0 + \sum_{j=1}^M \lambda_j \mu_j\) the exact entropy of \(f_M(x)\), where \(\lambda_j\) are obtained from (2.9) involving \(\{\mu_j = \mu_j(f_M)\}_{j=M+1}^{2M}\);

(ii) \(H^{\text{app}}[f_M] = (\lambda_0 + \Delta \lambda_0) + \sum_{j=1}^M (\lambda_j + \Delta \lambda_j) \mu_j\) the approximate entropy of \(f_M(x)\), where \(\lambda_j + \Delta \lambda_j\) are obtained from (2.9) involving \(\{\mu_j = \mu_j(f)\}_{j=M+1}^{2M}\) and \(\lambda_0 + \Delta \lambda_0\) from (2.10);

(iii) error \(\Delta H = H[f_M] - H^{\text{app}}[f_M]\).

From Theorem 2, as \(M \to \infty\), \(\{\mu_j(f_M) = \mu_j(f_M)\}_{j=M+1}^{2M}\) and then \(H[f_M] = H^{\text{app}}[f_M]\). From Theorem 1, as \(M \to \infty\), \(H[f_M] = H[f]\) and then both \(H[f_M] = H^{\text{app}}[f_M]\) and \(H[f] = H^{\text{app}}[f_M]\). Equivalently, \(H[f]\) is evaluated using directly \(\{\mu_j(f)\}_{j=1}^{2M}\). The computational cost amounts to solving a linear system (2.9) and a definite integral (2.10).

**Remark 2.3.** Relationships (2.9) and (2.10) are well known but unfortunately are used in an incorrect way by some authors. Such relationships have been considered as a short way to calculate Lagrange multipliers \(\{\lambda_j\}_{j=1}^M\) through the solution of a linear system avoiding the computational burden represented by (2.4). With such an improper procedure one replaces the sequence \(\{\mu_j(f)\}_{j=M+1}^{2M}\) (which will is known only after calculating \(\{\lambda_j\}_{j=1}^M\) through (2.4)) with \(\{\mu_j(f)\}_{j=M+1}^{2M}\). Nevertheless, for high \(M\) values (2.9) and (2.10) may be considered as a valuable tool to calculating \(\{\lambda_j\}_{j=1}^M\), solving a linear system and a definite integral, without solving nonlinear optimization problem (2.4). Indeed, the replacing \(\{\mu_j(f_M)\}_{j=M+1}^{2M}\) with \(\{\mu_j(f)\}_{j=M+1}^{2M}\) becomes meaningful as \(f_M(x) = f(x)\) a.e., in virtue of the entropy convergence theorem. At the end, the entropy \(H[f] = H[f_M] = \lambda_0 + \sum_{j=1}^M \lambda_j \mu_j\) can be found in terms of \(\{\mu_j(f)\}_{j=1}^{2M}\) only.

### 2.1. \(\Delta H\) estimate

In the special case when the convergence of \(H[f_M]\) to \(H[f]\) is fast, an estimate of \(\Delta H = H[f_M] - H^{\text{app}}[f_M]\) may be obtained as follows.

Let us consider (2.9). If \(\{\mu_j = \mu_j(f_M)\}_{j=M+1}^{2M}\) are involved, then (2.9) may be written in matrix form as \(\Delta_{2M} \cdot \lambda = b\), with

\[
b = \begin{bmatrix}
1 - 2\mu_1 \\
1 - 3\mu_2 \\
\vdots \\
1 - (M + 1)\mu_M
\end{bmatrix}, \quad \Delta_{2M} = \begin{bmatrix}
\mu_1 - \mu_2 & \cdots & \mu_M - \mu_{M+1} \\
\vdots & \ddots & \vdots \\
\mu_1 - \mu_{M+1} & \cdots & \mu_M - \mu_{2M}
\end{bmatrix} \text{diag}(1, 2, \ldots, M).
\]

If \(\{\mu_j = \mu_j(f_M)\}_{j=M+1}^{2M}\) are involved, then (2.9) may be written as \((\Delta_{2M} + \delta)(\lambda + \Delta \lambda) = b\), where the matrix \(\delta\) is obtained as a difference of matrices whose entries are given by \(\mu_j(f_M)\) and \(\mu_j(f)\), respectively \(\delta = \Delta_{2M}(\mu_j = \mu_j(f_M)) - \Delta_{2M}(\mu_j = \mu_j(f)), j = 1, \ldots, 2M\) and taking into account \(\mu_j(f_M) = \mu_j(f), j = 1, \ldots, M\). Then \(\Delta_{2M}\) is a Hankel matrix with on and lower antidiagonal entries given by \(\Delta_{2M} = \mu_j(f_M) - \mu_j(f), j = 1, \ldots, M\). From (2.14) it follows \(||\delta||_1 \leq M\sqrt{2[H[f_M] - H[f]]}\) and from \(\lambda = \Delta_{2M} \cdot \lambda = \Delta_{2M} + \delta\) it follows

\[
\Delta H = H[f_M] - H^{\text{app}}[f_M] = \lambda_0 + \sum_{j=1}^M \lambda_j \mu_j - (\lambda_0 + \Delta \lambda_0) - \sum_{j=1}^M (\lambda_j + \Delta \lambda_j) \mu_j = -\Delta \lambda_0 - \sum_{j=1}^M \Delta \lambda_j \mu_j
\]

and

\[
|\Delta H| = |H[f_M] - H^{\text{app}}[f_M]| \leq |\Delta \lambda_0| + \sum_{j=1}^M |\Delta \lambda_j| |\mu_j| \leq |\Delta \lambda_0| + \sum_{j=1}^M |\Delta \lambda_j| = |\Delta \lambda_0| + ||\lambda||_1.
\]

It remains to evaluate \(|\Delta \lambda_0|\) and \(||\lambda||_1\). If the convergence of \(H[f_M]\) to \(H[f]\) is fast, so that from (2.14) \(|\Delta \lambda_0| \leq |\Delta \lambda_0| ||\lambda||_1| + |\Delta \lambda_0| B |\lambda||_1| = |\Delta \lambda_0| M \sqrt{2[H[f_M] - H[f]]} \leq 1\) and as well as \((I + \Delta_{2M} \delta)^{-1} - \Delta_{2M} \delta\) may be assumed, then it holds

\[
||\Delta \lambda||_1 \leq ||\Delta \lambda + \Delta \lambda||_1 \leq ||(\Delta \lambda_0 + \Delta \lambda_0) - \lambda_0 - \lambda_0||_1 \leq \sum_{j=1}^M ||\Delta \lambda_j| |\mu_j| |_1 \leq ||\delta||_1 \leq |\Delta \lambda_0| + |\Delta \lambda_1| \leq |\Delta \lambda_0| + |\Delta \lambda_1| \leq |\Delta \lambda_0| + ||\lambda||_1, 
\]

(2.15)
For what $\Delta \lambda_0$ is concerned, from (2.13) we have
\[
\Delta \lambda_0 = (\lambda_0 + \Delta \lambda_0) - \lambda_0 = \ln \int_0^1 \exp \left( - \sum_{j=1}^M (\lambda_j + \Delta \lambda_j) x^j \right) dx
\]
\[
= \ln \int_0^1 \exp \left( - \sum_{j=1}^M \lambda_j x^j \right) \exp \left( - \sum_{j=1}^M \Delta \lambda_j x^j \right) dx
\]
\[
= \ln \left( 1 - \sum_{j=1}^M \Delta \lambda_j \mu_j \right) \simeq - \sum_{j=1}^M \Delta \lambda_j \mu_j,
\]
from which
\[
|\Delta \lambda_0| \leq \sum_{j=1}^M |\Delta \lambda_j| |\mu_j| \leq \|\Delta \lambda\|_1.
\]

Then from (2.15) one obtains
\[
|\Delta H| \leq |\Delta \lambda_0| + \|\Delta \lambda\|_1 \leq 2M^2 \sqrt{2(H[fM] - H[f])\|\Delta \lambda\|_1^2}\|b\|_1,
\]
the wanted bound for $\Delta H$.

### 3. Numerical results

In order to investigate the numerical consistency of the proposed procedure we apply it to several cases in which the target pdf is known and then any number of its moments can be calculated, as well as its entropy. Once assigned $\{\mu_j \equiv \mu(j)^{2M}\}$ for increasing $M$ values, $H_{\text{app}}[f_{\text{M}}]$ is obtained from (2.9) and (2.10). From (2.5) and Theorem 2, $H[f] = H[f_M] = H_{\text{app}}[f_{\text{M}}]$ holds true for large $M$ values. As $H_{\text{app}}[f_{\text{M}}]$ is obtained through a simplified procedure, which turns out true for high $M$ values only, we expect that, for some low $M$ values, $H_{\text{app}}[f_{\text{M}}] < H[f]$ or the sequence $\{H_{\text{app}}[f_{\text{M}}], M = 2, 3, \ldots\}$ fails to be monotonic decreasing without contradicting the MaxEnt procedure. Then the asymptotic value of $H_{\text{app}}[f_{\text{M}}]$ provides the wanted value $H[f]$.

1. $H_{\text{app}}[f_{\text{M}}] = \lambda_0 + \sum_{j=1}^M \lambda_j \mu_j$ with $\{\lambda_j\}_{j=1}^M$ and $\lambda_0$ given by (2.9) and (2.10) respectively, using $\{\mu_j \equiv \mu(j)^{2M}\}$;
2. $H[f_{\text{M}}]$ obtained from $f_{\text{M}}(x) = \exp \left( - \sum_{j=1}^M \lambda_j x^j \right)$, whose Lagrange multipliers $\{\lambda_j\}_{j=1}^M$ are calculated by (2.4);
3. exact entropy $H[f]$;
4. relative errors $\frac{|H_{\text{app}}[f_{\text{M}}] - H[f]|}{H[f]}$ and $\frac{|H_{\text{app}}[f_{\text{M}}] - H[f]|}{H[f]}$.

All the computations have been done by high numerical precision, using Maple.

### 3.1. Preconditioner of $\Delta 2M$

For what concerns the numerical solution of (2.9), let us consider the matrix
\[
\Delta 2M = \begin{bmatrix}
\mu_1 - \mu_2 & \cdots & \mu_M - \mu_{M+1} \\
\vdots & \ddots & \vdots \\
\mu_1 - \mu_{M+1} & \cdots & \mu_M - \mu_{2M}
\end{bmatrix} = \begin{bmatrix} 1 \cdots 1 \end{bmatrix} \text{diag}(\mu_1, \ldots, \mu_M) - \begin{bmatrix} \mu_2 & \cdots & \mu_{M+1} \\
\vdots & \ddots & \vdots \\
\mu_{M+1} & \cdots & \mu_{2M}\end{bmatrix} = \Delta 2M \quad (3.1)
\]
splitted in two matrices, $\Delta (1) 2M$ and $\Delta (2) 2M$; the last one is a Hankel matrix (for brevity $\Delta 2M$ maintains same name as the matrix involved in (2.9)). Now $\Delta 2M$ is a moments matrix relating Lagrange multipliers $\lambda$ with moments $\{\mu_j(f_{\text{M}})\}_{j=1}^M$ and it is ill-conditioned, with condition number comparable to that of $\Delta 2M$. This severe ill-conditioning motivates the search for a preconditioner for $\Delta 2M$ that can reduce the illness. For instance, a comparison between $\text{Cond}(\Delta 2M)$ and $\text{Cond}(\Delta (2) 2M)$ is illustrated in Fig. 1 in order to prove the conjecture that both matrices have a comparable condition number, since both are moments matrices (here the involved moments come from $f(x)$ which turns out used in next Example 4; the ratio $\text{Cond}(\Delta 2M)/\text{Cond}(\Delta (2) 2M)$ is reported, according to (3.2) below). Even if asymptotically $\text{Cond}(\Delta 2M)/\text{Cond}(\Delta (2) 2M)$ → 1, a good preconditioner for $\Delta 2M$ not necessarily is a good preconditioner for $\Delta 2M$.

It remains to estimate the condition number of $\Delta 2M$. A result in [10] (Theorem 3.2) states Hankel matrices $\Delta (2) 2M$ generated by moments of a strictly positive weight function with support the interval $[0, 1]$ asymptotically are conditioned in the same way as Hilbert matrices $H_M$ of the equal size, i.e.
moments of a density such that its moments are available in closed form involving some special functions such that. Then, in some sense, results in [10, Theorem 3.3] may be extended to non-negative weight functions with consequent improved estimate of $\lambda_j$, $j = 0, 1, \ldots, M - 1$ (21) for more details). By solving the preconditioned system higher $M$ values may be taken into account with consequent improved estimate of $H[f]$. A comparison between condition number of $\Delta_{2M}$ given by (2.9) and preconditioned matrix $C_M \Delta_{2M} C_M^{-1}$, for increasing $M$, is illustrated in Fig. 2 (some experiments with different $f(x)$ provide comparable results). Here the entries of $\Delta_{2M}$ are the moments of $f(x)$ used in Example 4. The improvement of conditioning is not trivial.

**Remark 3.1.** In [10, Theorem 3.2] strictly positive weight functions $f(x)$ with support $[0, 1]$ are considered, i.e., $\int_0^1 f(x) > 0$, $\forall x \in [0, 1]$. Such special weight functions are conditioned as same order Hilbert matrices. On the other hand in [15] it was proved each Hankel matrix $\Delta_{2M}$ generated by moments of MaxEnt density $f_M(x)$ has condition number satisfying $\text{Cond}(\Delta_{2M}) \geq 2^{4M-2}$. Taking into account Theorems 1 and 2 we may conclude that each Hankel matrix generated by moments of a density $f(x)$, with $\int_0^1 f(x) > 0$, $\forall x \in [0, 1]$ has condition number asymptotically satisfying $\text{Cond}(\Delta_{2M}) \geq 2^{4M-2}$. Then, in some sense, results in [10, Theorem 3.3] may be extended to non-negative weight functions $f(x) \geq 0$ and then $H_M$ may be used as a preconditioner in all next examples.

### 3.2. Numerical examples

All the below considered densities have support $[0, 1]$ and all next examples are borrowed from [20]. Each chosen pdf is such that its moments are available in closed form involving some special functions such that.
Example 1. Let us consider the pdf

\[ f(x) = A e^{-x^n}, \quad A = \frac{n}{\Gamma\left(\frac{n}{n} \Gamma_{\text{inc}}\left(1, \frac{1}{n}\right) \right)} \quad n \in \mathbb{R}^+. \]

The change of variable \( x^n = t \) allows us to calculate the above normalizing constant \( A \), moments and entropy in terms of Incomplete Gamma function

\[ \mu_j = \frac{\Gamma\left(\frac{n+1}{n} \Gamma_{\text{inc}}\left(1, \frac{1}{n}\right) \right)}{\Gamma\left(\frac{n}{n} \Gamma_{\text{inc}}\left(1, \frac{1}{n}\right) \right)}, \quad j = 0, 1, \ldots \]

(see also [20, 8.326.10, p. 337]) and entropy

\[ H[f] = -\ln \left[ \frac{n}{\Gamma\left(\frac{n}{n} \Gamma_{\text{inc}}\left(1, \frac{1}{n}\right) \right)} + \frac{\Gamma\left(\frac{n+1}{n} \Gamma_{\text{inc}}\left(1, \frac{1}{n}\right) \right)}{\Gamma\left(\frac{n}{n} \Gamma_{\text{inc}}\left(1, \frac{1}{n}\right) \right)} \right]. \]

The density \( f(x) \) is MaxEnt density with characterizing moment \( E(X^n) \). Then, whenever \( n \in \mathbb{N} \), once assigned \( (\mu_j)_{j=0}^n \) MaxEnt \( f_n(x) = \exp\left(\sum_{j=0}^n \mu_j x^j\right) \) coincides with \( f(x) \) and \( H[f_n] = H[f] \).

The example, where \( n = 15 \) is assumed and then \( H[f] \approx -0.00980278980538 \), has as a goal to prove experimentally (1.1) which states higher moments may be well approximated by a sequence of lower moments. Then the information content is spread among lower moments (see Fig. 3). Since \( H[f] \approx 0 \) the bound (1.2) becomes meaningful. The latter provides the estimate \( H[f] \approx 0.001987 \).

Example 2. Let us consider a Beta density

\[ f(x) = \frac{1}{B(p, q)} x^{p-1} (1-x)^{q-1} \quad \text{with} \quad p = 4, \quad q = 2. \]

![Figure 3](image-url) - (a) \( H^{\text{approx}}[f_{\text{inc}}] (\cdot) \) and exact entropy \( H[f] \) (dashed line); (b) Relative errors \( \left| \frac{H^{\text{approx}}[f_{\text{inc}}] - H[f]}{H[f]} \right| (\cdot) \) and \( \left| \frac{H^{\text{approx}}[f_{\text{inc}}] - H[f]}{H[f]} \right| (\cdot) \) for Example 1.
with moments [20, 3.199, p. 318] obtained directly in terms of Beta function

$$\mu_j = \frac{\Gamma(p + q)\Gamma(p + j)}{\Gamma(p + q + j)\Gamma(p)}, \quad j = 0, 1, \ldots$$

and Shannon entropy given in terms of Beta and Digamma function

$$H[f] = \ln \beta(p, q) - (p - 1)(\psi(p) - \psi(p + q)) - (q - 1)(\psi(q) - \psi(p + q)) = \ln 20 + 79/30 \approx -0.3623989402206575.$$  

Now $H^{\text{app}}[f_M], H[\text{f_M}]$ for increasing $M$ and $H[f]$ are reported in Fig. 4. Here $H[f_M], M = 2, \ldots, 8$ is reported. Although $H^{\text{app}}[f_M]$ be calculated through an approximate procedure, its value is more accurate than $H[f_M]$. This seems reasonable: indeed $H^{\text{app}}[f_M]$ involves $2M$ moments, whilst $H[f_M]$ involves the first $M$ moments.

**Example 3.** Let us consider the density

$$f(x) = -\frac{2}{\pi \ln(2)} \frac{\ln(x)}{\sqrt{1 - x^2}}.$$  

Starting from $l(a) = \int_0^1 \frac{x^n}{\sqrt{1 - x^2}} \, dx$, from which $\frac{d^2}{dx^2} l(a) = \int_0^1 x^n \, dx \frac{\ln(x)}{\sqrt{1 - x^2}} \, dx$. The change of variable $x = \sin \theta$ leads to $l(a) = \int_0^{\pi/2} \sin^n \theta \, d\theta$, so that the wanted integral is given in term of Beta function derivatives. When $a$ takes integer values, after some algebra one obtains moments (see also [20], 4.241–2, p. 538)

$$\mu_0 = 1,$$

$$\mu_{2n} = -\frac{2}{\pi \ln(2)} \frac{2n\pi}{2^{2n} - 1} \left(-\ln(2) + \sum_{k=1}^{2n} \frac{(-1)^{k-1}}{k}\right), \quad n = 1, 2, \ldots,$$

$$\mu_{2n+1} = -\frac{2}{\pi \ln(2)} \frac{2^{2n}}{(n+1)!} \left(\ln(2) + \sum_{k=1}^{2n+1} \frac{(-1)^{k}}{k}\right), \quad n = 0, 1, 2, \ldots,$$

whilst the entropy, obtained numerically, is

$$H[f] \approx -0.3179752927304.$$  

Numerical results are reported in Fig. 5 and all the considerations are analogous to **Example 2**.
Example 4. Let us consider the density
\[ f(x) = -\frac{8}{\pi(1 + 2\ln(2))} \ln(x)\sqrt{1 - x^2}. \]
. Moments and entropy are obtained with procedure similar to previous example ([20], 4.241.3–4, p. 538)

\[
\begin{align*}
\mu_0 &= 1, \\
\mu_{2n} &= -\frac{8}{\pi(1 + 2\ln(2))} \frac{\binom{2n}{n}}{2^{2n+1}(n+1)} \left( \ln(2) + \frac{1}{2n+2} + \sum_{k=1}^{2n} \frac{(-1)^k}{k} \right), \quad n = 1, 2, \ldots, \\
\mu_{2n+1} &= -\frac{8}{\pi(1 + 2\ln(2))} \frac{2^{2n+1}}{(n+1)(n+2)} \left( \ln(2) - \frac{1}{2n+3} + \sum_{k=1}^{2n+1} \frac{(-1)^k}{k} \right), \quad n = 0, 1, 2, \ldots
\end{align*}
\]

and

\[ H[f] \approx -0.508931221953. \]

Numerical results are reported in Fig. 6 and all the considerations are analogous to Example 2.

Example 5. Let us consider the truncated Cauchy density
\[ f(x) = \frac{4}{\pi(1 + x^2)}. \]

Here odd and even \( \mu_j \) are calculated through the stable recursive relationship

\[
\mu_j = \frac{4}{\pi} \left( \frac{1}{j - 1} - \mu_{j-2} \right), \quad j \geq 2, \quad \mu_0 = 1, \quad \mu_1 = \frac{2}{\pi} \ln 2.
\]

(closed formulas are in ([20], 2.146.1, p. 76)) whilst \( H[f] \approx -0.0215137302739 \). \( f(x) \) may be written as \( f(x) = e^{\ln \left( \frac{1}{\pi(1 + x^2)} \right)} \approx e^{\ln 2} \cdot x^2 \cdot x^4 \cdot 2 \cdot x^6 \cdot 3 \ldots \) and then \( f(x) \) may be considered as MaxEnt distribution constrained by integer moments. It is reasonable expect few integer moments are able to catch the information content of \( f(x) \). Numerical results are reported.
in Fig. 7; entropy decreasing is fast as $M$ increases and the conjecture about the information content of $f(x)$ is proved. Here MaxEnt distributions are calculated until $M = 6$ before incurring in numerical instability. Since $H[f]$ is close to zero the bound (1.2) becomes meaningful. The latter provides the estimate $H[f] \leq -0.003338$.
Example 6. In this example we consider a pdf $f(x)$ so that the procedure to estimate $f(x)$ from $\{\mu_j\}_{j=0}^\infty$ completely fails. Let us consider the highly oscillating density

$$f(x) = A \sin^2(\pi \omega x), \quad \omega \gg 1,$$

with $A = \frac{1}{\pi^2}$. Taking into account well known relationships

$$\sin^2(\pi \omega x) = \frac{1}{2} (1 - \cos(2\pi \omega)) \quad \text{and} \quad \lim_{\omega \to \infty} \int_0^1 f(x) \cos(2\pi \omega x) dx = 0,$$

for each piecewise continuous function $f(x)$, we have as $\omega \to \infty$

$$A = 2; \quad \mu_n = A \int_0^1 x^n \sin^2(\pi \omega x) dx = \frac{A}{2} \int_0^1 x^n dx - \frac{A}{2} \int_0^1 x^n \cos(2\pi \omega x) dx = \frac{1}{n+1}$$

(which coincides with the moments of the uniform distribution); and, after some algebra,

$$H[f] = -\int_0^1 A \sin^2(\pi \omega x) \ln(A \sin^2(\pi \omega x)) dx = -1 + \ln 2 \approx -0.30685281944.$$

Then $H[f] \approx -0.30685281944$, while $H^{app}[f_M] = H[f_M] = 0, \forall M$, since $\mu_n = \frac{1}{n+1}$ implies that MaxEnt density $f_M(x) = 1$, the uniform distribution and from (2.9) and (2.10) $\lambda_1 = \cdots = \lambda_M = 0, \lambda_0 = 1$ holds. Then, if $f(x)$ differs from a “reasonable density”, the outlined procedure may completely fail in evaluating $H[f]$ from $\{\mu_j(f)\}_{j=0}^\infty$. If several moments are unable to squeeze the relevant part of information about $f(x)$, it should mean that moments represent an unsuitable tool. As a consequence, we should turn our attention to more proper expected values.

Remark 3.2. Numerical results highlight

1. $H^{app}[f_M] \geq H[f], \forall M$;  
2. the sequence $\{H^{app}[f_M]\}$ is monotonic decreasing.

Hence we prove the following assumptions through an heuristic reasoning:

(a) as $M \to \infty$, $H^{app}[f_M] = H[f_M] \geq H[f]$, being $\mu_j(f_M) \equiv \mu_j(f), j = 1, \ldots, 2M$;  
(b) for each fixed $M$, $H^{app}[f_M]$ takes into account $\mu_j(f), j = 1, \ldots, 2M$, while $H^{app}[f_{M-1}]$ considers $\mu_j(f), j = 1, \ldots, 2M - 2$.

Then, according to physical meaning of $H^{app}[f_M]$, it happens that $H^{app}[f_{M-1}] > H^{app}[f_M]$ proving that the sequence $\{H^{app}[f_M]\}$ is monotonic decreasing. Combining the previous two facts, we obtain the following criterion for the choice of the number $M$ of moments that yield the best approximation $H^{app}[f_M]$ before incurring in numerical instability (equivalently, the sequence $\{H^{app}[f_M]\}$ fails to be monotonic decreasing):

$$M \equiv \min_j H^{app}[f_M].$$

The failure comes from an insufficient machine precision.

4. Conclusions and discussion

The problem of evaluating Shannon entropy of a distribution, whose only the sequence of integer moments is given, is considered. Shannon entropy is directly given in terms of integer moments without solving the underlying moment problem. Known results about both entropy-convergence of MaxEnt approximation and Kullback–Leibler distance allowed us to reach the goal. Ill-conditioning issues affecting the linear system to be solved may be mitigated by a proper pre-conditioning technique with Hilbert matrix as a preconditioner. Numerical results prove the moments are a suitable tool to gather the information content of a distribution, even if an accurate entropy estimate requires the involving a high number of moments. The common opinion diffused in literature that the information content of a distribution is spread among the sequence of infinite moments is now supported by numerical evidence. If, in the approximation framework, the discrepancy between underlying density and its approximation is measured by Kullback–Leibler distance then the numerical tests allow us the following conclusions.

(a) a high number of moments have to be involved with consequent issues of ill-conditioning;  
(b) more suitable expected values, as fractional moments $E[X^k] = \int_0^1 x^k f(x) dx, x \in \mathbb{R}$, have to be involved for an accurate recovering. Two of the authors of this paper have suggested the use of fractional moments in some past papers [22,23]. Numerical results illustrated above support that theoretical conjecture.
(c) As a further alternative to integer moments in [24] the use of shifted Chebyshev polynomials $T_n(x) = T_n(2x - 1)$, with $T_n(x) = \cos[n\cos^{-1}(x)]$ is suggested. More accurate results obtained, than integer moments, can be explained by taking into account the superior minimax property of the Chebyshev polynomials and the consequent analytical form of the MaxEnt solution $f_M(x) = \exp\left(\frac{-\sum_{j=0}^{M-1} \lambda_j x^j}{\lambda_0}\right)$ for Chebyshev moments. Indeed, in presence of the term $\lambda_j x^j$ in the MaxEnt solution (2.1) for integer moments makes it very difficult to exploit information from the higher order moments in the interval $[0,1]$. Such an issue does not appear in the formulation of the problem via Chebyshev moments owing to the bounded nature of Chebyshev polynomials and more regular sampling of the interval, and consequently provides a way to incorporate systematically the information from the higher moments in the reconstruction process.

The procedure outlined here above for the Hausdorff case might be likewise extended to Stieltjes and Hamburger cases. Indeed, in Stieltjes case (in Hamburger case the considerations are similar; even values of $M$ are requested) higher moments $\mu_j(f_M)$, $j \geq M$ may be found in terms of $\left\{\mu_j\right\}_{j=1}^M$ and $\left\{\lambda_j\right\}_{j=1}^M$ integrating by parts (2.2) (here $D \equiv \mathbb{R}^+$) with the following result

$$\sum_{k=1}^{M} k\lambda_k \mu_{k-j} = (1+j)\mu_j, \quad j = 0, \ldots, M-1$$

and

$$\lambda_0 = \ln \int_{0}^{\infty} \exp\left(-\sum_{k=1}^{M} k\lambda_k x^k\right) dx.$$  

In analogy with Hausdorff case, MaxEnt density $f_M$ converges in entropy to $f$ [25] and then in directed divergence, so that $\lim_{M \to \infty} f_M(x) = f(x)$ almost everywhere. Then, starting from high values of $M$ we may assume $\mu_j(f_M) = \mu_j(f)$, for each $j, k$, with $j + k \geq M$, where $\mu_j(f)$ are known quantities. (4.1) may be considered a linear system which provides $(\lambda_1, \ldots, \lambda_M)$. The linear system (4.1) admits solution, being the involved matrix a Hankel matrix. Nevertheless two main issues are still open.

1. A suitable preconditioner for Hankel matrices coming from Stieltjes (or Hamburger) moment problem is lacking in literature. Then high $M$ values cannot be used in (4.1).
2. How to ensure $\lambda_M \neq 0$ in (4.1) to ensure integrability in (4.2).

The last drawback may be solved as follows. An usual practice amounts to truncate the support $\mathbb{R}^+$, so that Stieltjes moment problem is solved within a proper large enough interval $[0, b]$ one might be chosen according to Chebyshev inequality

$$P(\lvert X - \mu_1 \rvert \geq c) \leq \frac{\mu_1 + c}{\mu_1 - c},$$

changing the original Stieltjes problem into Hausdorff moment problem (in such a case $\lambda_M$ may assume any real value). In MaxEnt setup the latter admits a solution for each sequence $\left\{\lambda_j\right\}_{j=1}^M$ inner to moment space [26, Theorem 4]. On the contrary, in the Stieltjes case the moment space of Maxent $f_M$ is a proper subset of the moment space and then, for an arbitrary set $\left\{\mu_j\right\}_{j=1}^M$, the existence of $f_M$ is based only on the numerical evidence [27, Theorem 3]. Then we might be led to the unpleasant fact that a MaxEnt solution always exists, although the original Stieltjes moment problem doesn’t admit any solution. In conclusion, the extension to Stieltjes (or Hamburger) case of the illustrated technique to calculating the entropy directly from the moments sequence is not obvious.

References