A NEW BOUND FOR DISCRETE DISTRIBUTIONS BASED ON MAXIMUM ENTROPY

Henryk Gzyl*, Pier Luigi Novi Inverardi† and Aldo Tagliani†

*USB and IESA - Caracas (Venezuela)
†Dept. of Computer and Management Sciences - University of Trento - 38100 Trento (Italy)

Abstract.
In this work we re-examine some classical bounds for nonnegative integer-valued random variables by means of information theoretic or maxentropic techniques using fractional moments as constraints. The new bound is able to capture optimally all the information content provided by the sequence of given moments or by the moment generating function, summarized by few fractional moments. The bound improvement is not trivial.

Keywords: Distribution bounds, Entropy, Fractional moments, Moments, Tail probability

SOME KNOWN BOUNDS FOR DISCRETE PROBABILITY DISTRIBUTIONS

Consider a non negative integer-valued r.v. X with distribution function $F(x)$. We often need to compute the survival probability

$$P(X \geq t) = 1 - F(t) = \int_t^\infty dF(x) \tag{1}$$

or $F(x)$ itself or expected values. For many cases of interest (1) is not explicitly given in closed form, so that we must be satisfied with providing upper bounds of (1). Classically, three candidates are most used as an upper bound of (1):

1) the well known Chernoff bound $C(t)$ ([2]) defined by

$$C(t) = \inf_{s \geq 0} M(s)\, e^{-st} \tag{2}$$

with

$$M(s) = \int_0^\infty e^{sx} dF(x),\, s \in I_\delta,\, \delta > 0$$

where I_δ is some complete neighborhood of the origin so that $M(s)$ is the usual moment generating function (*mgf*). For reasons which will be clear later, we introduce the function $M^*(s) = M(s)$, $s \in (-\infty, 0]$: note that $M^*(s)$ is not defined on a complete neighborhood of the origin and it has not be confused with the *fgm*.

2) the moment bound ([15]),

$$M_{mom}(t) = \inf_{n \geq 0} \frac{\mathbb{E}(X^n)}{t^n} \tag{3}$$

3) the factorial moment bound

$$\mathscr{F}(t) = \inf_{0 \leq n \leq t} \frac{\mathbb{E}\left(X(X-1)\cdots(X-n)\right)}{t(t-1)\cdots(t-n)} \tag{4}$$

These three bounds come from the Markov inequality. [15] showed that $M_{mom}(t)$ is better than $C(t)$, i.e. $P(X \geq t) < M_{mom}(t) < C(t)$. [13] showed that $\mathscr{F}(t)$ is better than $C(t)$, i.e. $P(X \geq t) < \mathscr{F}(t) < C(t)$.

Closely related to survival probability estimation is the following:

4) a classical bound ([1]) states that, if $F(x)$ and $G(x)$ are two distribution functions sharing the first $2Q$ moments $\mu_j = \int_0^\infty x^j dF(x) = \int_0^\infty x^j dG(x)$, $j = 1, 2, \ldots, 2Q$ then

$$\mid [1 - F(x)] - [1 - G(x)] \mid = \mid F(x) - G(x) \mid \leq \omega_Q(x) \tag{5}$$

where the window function $\omega_Q(x) = \left[V_Q'(x)\Delta_Q^{-1}V_Q(x)\right]^{-1}$ with

$$\Delta_Q = \begin{bmatrix} \mu_0 & \cdots & \mu_Q \\ \vdots & \vdots & \vdots \\ \mu_Q & \cdots & \mu_{2Q} \end{bmatrix}$$

the Hankel matrix and $V_Q(x) = [1, x, \ldots, x^Q]'$ is the so-called power vector.

[11] showed that (5) gives relatively sharp information about the tail of the distribution but, not too much else as consequence of the structure of $\omega_Q(x)$ which goes to zero at the rate x^{-2Q} as $x \to \infty$.

The above classical bounds concerning the distribution function $F(x)$ are given in terms of integer moments or in terms of moment generating function (*mgf*); hence, they exploit only partially the information contained in the data and for this reason these bounds are not very tight. Nevertheless, they are easily calculated using that data.

Fractional moments given by $\mathbb{E}(X^\alpha)$, $\alpha \in \mathbb{R}^+$ are definitely better than integer moments for recovering a probability distribution and related quantities via Maximum Entropy setup for several reasons; in particular, there is a result due to Lin ([10]) which states the characterization of a distribution through its fractional moments

Theorem 1 (Lin (1992)) *A positive r.v. X is uniquely characterized by an infinite sequence of positive fractional moments $\{\mathbb{E}(X^{\alpha_j})\}_{j=1}^\infty$ with distinct exponents $\alpha_j \in (0, \alpha^*)$, $\mathbb{E}(X^{\alpha^*}) < \infty$, for some $\alpha^* > 0$.*

and the Maximum Entropy *pmf* P_M recovered involving fractional moments converges in entropy to the true *pmf* P ([14]). This implies the convergence in directed divergence of P_M to P which implies convergence in L_1 norm of P_M to P ([9]) and, hence, convergence in distribution of bounded functions of X evaluated on P_M to the true value. This last result means that if we are interested in approximating some characteristic constants of a discrete distribution (think to expected values, probabilities or other) the equivalent counterparts evaluated on P_M are as close as we like to the true values and the closeness depends on the (increasing) value of M.

But, as a counterbalance, they are not always easy to evaluate. Traditionally the *mgf* of a random variable X is used to generate positive integer moments of X. But it is clear that the *mgf* also contains a wealth of knowledge about arbitrary real moments and hence, on fractional moments. Taking this into account, to obtain fractional moments, [5] exploit some properties of the *mgf* and its fractional derivatives; [8], in addition to *mgf*, considers the knowledge of a set of integer moments which can be obtained by proper integration of the *mgf* on a contour $\mathscr{C}$ of the complex plane.
Several scenarios will be analyzed, depending on the available information. This latter is assumed given by a finite or infinite sequence of moments and/or by the *mgf*.

The case $X \geq 0$ with determinate moment problem non admitting *mgf*

Let $X = \{x_0, x_1, \ldots\}$ be a discrete r.v. with probability mass function (*pmf*) $P = \{p_0, p_1, \ldots\}$ whose integer moments (im) $\mu_j = \sum_{k=0}^{\infty} x_k^j p_k$, $j = 1, 2, \ldots,$ are assigned. Uniquely in terms of moments the non existence of *mgf*, given in this case by $M(s) = \sum_{j=0}^{\infty} e^{s x_j} p_j$, entails

$$\limsup_{j\to\infty} \left(\frac{\mu_j}{j!}\right)^{\frac{1}{j}} = +\infty.$$

while moment problem determinacy entails ([12])

$$\lim_{n\to\infty} \rho_n^{(0)} \cdot \rho_n^{(1)} = 0$$

where

$$\rho_n^{(0)} = \frac{\begin{vmatrix} \mu_0 & \cdots & \mu_n \\ \cdots & \cdots & \cdots \\ \mu_n & \cdots & \mu_{2n} \end{vmatrix}}{\begin{vmatrix} \mu_2 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots \\ \mu_{n+1} & \cdots & \mu_{2n} \end{vmatrix}} \text{ and } \rho_n^{(1)} = \frac{\begin{vmatrix} \mu_1 & \cdots & \mu_{n+1} \\ \cdots & \cdots & \cdots \\ \mu_{n+1} & \cdots & \mu_{2n+1} \end{vmatrix}}{\begin{vmatrix} \mu_3 & \cdots & \mu_{n+2} \\ \cdots & \cdots & \cdots \\ \mu_{n+2} & \cdots & \mu_{2n+1} \end{vmatrix}}.$$

Next *ME* approximant $P_M^{(\mathrm{im})} = \{p_0^{(\mathrm{im})}, p_1^{(\mathrm{im})}, \ldots\}$ ([7]) of P, constrained by μ_j, $j = 1, ..., M$ is considered. Here

$$p_i^{(\mathrm{im})} = \exp\left(-\sum_{j=0}^{M} \lambda_j x_i^j\right)$$

with λ_j, $j = 0, 1, \ldots, M$, $\lambda_M \geq 0$ Lagrange multipliers. The constraints $\{\mu_j\}_{j=0}^{M}$ determine uniquely λ_j and hence $P_M^{(\mathrm{im})}$. If the underlying moment problem is determinate,

[16] proved that $P_M^{(\mathrm{im})}$ converges in entropy to P, that is $\lim_{M\to\infty} H[P_M^{(\mathrm{im})}] = H[P]$, where $H[P_M^{(\mathrm{im})}]$ and $H[P]$ denote the Shannon-entropy of $P_M^{(\mathrm{im})}$ and P respectively, with $H[P] = -\sum_{j=0}^{\infty} p_j \ln p_j$ and similarly $H[P_M^{(\mathrm{im})}]$.
Entropy convergence entails convergence in variation and then in distribution. Indeed, keeping in mind that $P_M^{(\mathrm{im})}$ and P have same first M moments, combining the following well known relationship

$$H[P_M^{(\mathrm{im})}] - H[P] = \sum_{j=0}^{\infty} p_j \ln \frac{p_j}{p_j^{(\mathrm{im})}}.$$

and the inequality [4]

$$\sum_{j=0}^{\infty} p_j \ln \frac{p_j}{p_j^{(\mathrm{im})}} \geq \frac{1}{2\ln 2} \left(\sum_{j=0}^{\infty} \mid p_j - p_j^{(\mathrm{im})} \mid \right)^2$$

we have

$$\begin{aligned} \mid F_M^{(\mathrm{im})}(x) - F(x) \mid &= \mid \sum_{j\leq x} \left(p_j^{(\mathrm{im})} - p_j \right) \mid \\ &\leq \sum_{j\leq x} \mid p_j^{(\mathrm{im})} - p_j \mid \\ &\leq \sum_{j=0}^{\infty} \mid p_j^{(\mathrm{im})} - p_j \mid \\ &\leq \sqrt{2\ln 2 \left(H[P_M^{(\mathrm{im})}] - H[P] \right)} \end{aligned} \tag{6}$$

The righthand term is the required uniform bound to be compared with (5). In (6) $H[P_M^{(\mathrm{im})}]$ may be calculated, while, in general, $H[P]$ may be efficiently estimated from the sequence $H[P_j^{(\mathrm{im})}]$, $j = 1, 2, \ldots, M$ through a proper convergence accelerating process (Aitken Δ^2-method, for instance)).

The case $X \geq 0$ where both $\{\mu_j\}_{j=1}^K$ and $M^*(s)$ are known

The knowledge of $\{\mu_j\}_{j=1}^K$ and $M^*(s)$, $s \leq 0$ allows us to obtain fractional moments $\mathbb{E}(X^\alpha) = \sum_{j=1}^{\infty} x_j^\alpha p_j$, $0 < \alpha < K$, through the following formula due to Klar ([8])

$$\mathbb{E}(X^{r+N-1}) = (-1)^N \frac{\prod_{j=0}^{N-1}(r+j)}{\Gamma(1-r)} \int_0^{\infty} s^{-r-N} \left[M(-s) - \sum_{j=0}^{N-1} (-1)^j \frac{\mu_j s^j}{j!} \right] ds \tag{7}$$

with $r \in (0,1)$, $N = 1,2,...,K$ and $\alpha = r + N - 1 \in (0,N)$. Now,

a) for $N = 1$ the right hand side of (7) involves only $M^*(s)$: this is enough to obtain infinite fractional moments with exponents in $(0,1)$ and, via Lin's theorem, they are able to characterize the distribution. In this case, (7) reduces to that given by [5];

b) for $N > 1$ the right hand side of (7) involves both $M^*(s)$ and a set of N integer moments; infinite fractional moments with exponents in $(0,N)$ may be obtained from (7) and, via Lin's theorem, they are able to characterize the distribution again. In this sense, (7) may be also seen as a generalization of the Cressie and Borkent result.

It is important to note that the two sides of (7) are equivalent in information about the distribution; but, the fractional moments are able to condense more effectively the same information contained in $M^*(s)$ and in the set of N integer moments.
Next the *ME* approximant $P_M^{(\mathrm{fm})} = \{p_0^{(\mathrm{fm})}, p_1^{(\mathrm{fm})}, \dots\}$ of P ([7]), constrained by $\{\mathbb{E}(X^{\alpha})\}_{j=0}^{M}$, $\alpha_0 = 0, 0 < \alpha_j < K$, K arbitrarily fixed with $\mathbb{E}(X^K) < +\infty$ according to what Lin's characterization theorem ([10]) says, is considered where

$$p_i^{(\mathrm{fm})} = \exp\left(-\sum_{j=0}^{M} \lambda_j x_i^{\alpha_j}\right) \tag{8}$$

with λ_j, $j = 0,1,\dots,M$, $\lambda_M \geq 0$ Lagrange multipliers. [14] proved that, if

$$\alpha_j = \Delta\alpha\, j,\ \ j = 0,1,\dots,M, \Delta\alpha = \frac{K}{M}$$

where $\mathbb{E}(X^K) < +\infty$, then $P_M^{(\mathrm{fm})}$ converges in entropy to P, that is

$$\lim_{M\to\infty} H[P_M^{(\mathrm{fm})}] = H[f] \tag{9}$$

Such a result, joined with $H[P_M^{(\mathrm{fm})}] \geq H[f], \forall M$, allowed the useful choice of $\{\alpha_j\}_{j=1}^{M}$ with $0 < \alpha_j \leq K$ according to the following criterion

$$\{\alpha_j\}_{j=1}^{M}:\ \ H[P_M^{(\mathrm{fm})}] = \text{minimum}. \tag{10}$$

For the convergence in entropy, only $0 < \alpha_j \leq K, \forall j$ is required, no matter regarding the value of K; of course, smaller K slower the convergence in entropy of $P_M^{(\mathrm{fm})}$ to P. Unlike from the integer moments setup where the optimal moment sequence $\{\mu_1, \mu_2, \dots, \mu_{M+1}\}$ is obtained from $\{\mu_1, \mu_2, \dots, \mu_M\}$ just adding μ_{M+1}, the two optimal sequences $\{\alpha_1^{(M)}, \alpha_2^{(M)}, \dots, \alpha_M^{(M)}\}$ and $\{\alpha_1^{(M+1)}, \alpha_2^{(M+1)}, \dots, \alpha_{M+1}^{(M+1)}\}$, satisfying (10) with M and $M+1$ respectively, are completely disconnected in the fractional moments setup. In numerical experiments the corresponding *ME pmf* $P_M^{(\mathrm{fm})}$ has entropy $H[P_M^{(\mathrm{fm})}] \simeq H[P]$ starting from moderate values of M. From (10) the convergence of

$H[P_M^{(\mathrm{fm})}]$ to $H[P]$, for increasing M, is evidently faster than the convergence of $H[P_M^{(\mathrm{im})}]$ to $H[P]$, if the underlying moment problem is determinate, i.e.

$$H[P_M^{(\mathrm{fm})}] - H[P] < H[P_M^{(\mathrm{im})}] - H[P], \forall M. \tag{11}$$

The chain of inequalities similar to (6), provides us with

$$\mid F_M^{(\mathrm{fm})}(x) - F(x) \mid \leq \sqrt{2\ln 2\left(H[P_M^{(\mathrm{fm})}] - H[P]\right)}. \tag{12}$$

The bound (12) is sharper than (6). Numerical evidence or a convergence accelerating process, proves that

$$H[P_M^{(\mathrm{fm})}] \simeq H[P] \tag{13}$$

even for moderate values of M. By combining (6) and (13) we have the testable uniform bound

$$\mid F_M^{(\mathrm{im})}(x) - F(x) \mid \leq \sqrt{2\ln 2\left(H[P_M^{(\mathrm{im})}] - H[P_M^{(\mathrm{fm})}]\right)} \tag{14}$$

i.e. the upper bound is obtained through two different procedures, having different and comparable accuracy. Probably the bound (14) is sharper than (5) in the central part of the distribution and, vice versa, (5) is much more sharp than (14) (as well (12)) in the tail. Combining (5) and (14) (or (12)) we have a sharper upper bound, valid for $x \geq 0$ and M which guarantees (13)

$$\mid F_{2M}^{(\mathrm{im})}(x) - F(x) \mid \leq \min\{\omega_Q(x), \sqrt{2\ln 2\left(H[P_{2M}^{(\mathrm{im})}] - H[P_{2M}^{(\mathrm{fm})}]\right)}\} \tag{15}$$

$$\mid F_{2M}^{(\mathrm{fm})}(x) - F(x) \mid \leq \min\{\omega_Q(x), \sqrt{2\ln 2\left(H[P_{2M}^{(\mathrm{fm})}] - H^{acc}[P]\right)}\} \tag{16}$$

where $H^{acc}[P]$ is obtained from $\{H[P_j^{(\mathrm{fm})}]\}_{j=1}^{2M}$ through a convergence accelerating process, so that $H[P] \simeq H^{acc}[P]$ may be assumed. Here the maximum value allowed of Q stems from the number of given moments or from numerical stability requirements.

The case $X \geq 0$ with $\{\mu_j\}_{j=1}^{\infty}$ assigned and existing *mgf*

Let us now assume that we know $\{\mu_j\}_{j=1}^{\infty}$, and we also assume that

$$\limsup_{j\to\infty} \left(\frac{\mu_j}{j!}\right)^{\frac{1}{j}} = \frac{1}{R}, \quad \text{finite.}$$

Then $M(t) = \sum_{j=0}^{\infty} \frac{\mu_j t^j}{j!}$, $-R \leq t < R$, holds. Since our first goal is to calculate $\mathbb{E}(X^{\alpha})$ from $\{\mu_j\}_{j=1}^{\infty}$ through (7), it remains to determinate $M(t)$ on $(-\infty, -R]$. The following procedure is adopted. The underlying *mgf* $M(t)$ is such that $M(-t)$ is a completely monotonic function on $(-R,+\infty)$, i.e. $(-1)^j M^{(j)}(-t) > 0, \forall t > -R,\ j = 0,1,\ldots$; then $M(-t)$ may be uniformly approximated on $t \in [0,\infty)$ by the following exponential sum ([6])

$$M(-t) \simeq Y_n(-t) = \sum_{j=1}^{n} a_j e^{-\lambda_j t} \tag{17}$$

having parameters satisfying the constraints $0 \le \lambda_1 < \lambda_2 < \ldots < \lambda_n$, $a_i \ge 0$, $i = 1,\ldots,n$. Now, if $Y_n(-t)$ interpolates $M(-t)$ at the $2n$ equally spaced points $t_j \in [0,R]$, $j = 1,\ldots,2n$, then Prony's method may be invoked to calculate the parameters a_j's and λ_j's. Being $M(-t)$ a completely monotonic function with $M(-\infty) = 0$ then Prony's method guarantees $a_j \ge 0, \forall j$ and $0 \le \lambda_1 < \lambda_2 < \cdots < \lambda_n$ ([6], Thm. 3) so that $Y_n(-t)$ turns to be completely monotonic function too and asymptotically decreasing to zero. Then for practical purposes, $M(-t) \simeq Y_n(-t)$, $t \ge R$. Finally, by replacing $M(-t)$ with $Y_n(-t)$, $t \ge R$, fractional moments $\mathbb{E}(X^\alpha)$ are obtained by a slightly modification of (7) (Klar's formula)

$$\begin{aligned}\mathbb{E}(X^{r+N-1}) = (-1)^N \prod_{j=0}^{N-1} \frac{(r+j)}{\Gamma(1-r)} \Bigg(\int_0^R s^{-r-N} \Big[M(-s) - \sum_{j=0}^{N-1} (-1)^j \frac{\mu_j s^j}{j!} \Big] ds + \\ + \int_R^\infty s^{-r-N} \Big[Y_n(-s) - \sum_{j=0}^{N-1} (-1)^j \frac{\mu_j s^j}{j!} \Big] ds \Bigg)\end{aligned} \tag{18}$$

Next according to *ME* procedure the approximant *pmf* $P_M^{(\mathrm{fm})} = \{p_0^{(\mathrm{fm})}, p_1^{(\mathrm{fm})}, \ldots\}$ constrained by $\{\mathbb{E}(X^{\alpha_j}\}_{j=0}^{M}$, $\alpha_0 = 0$ is obtained ([7]) with $p_i^{(\mathrm{fm})}$ given by (8). The choice of $\{\alpha_j\}_{j=1}^{M}$ is similar to the one adopted in (9) as well as the entropy convergence.

The case $X \ge 0$ with $M(t), t \in (-\infty, R)$ known, R finite or infinite

This case is an extension of the previously analyzed case. By repeated differentiation of $M(t)$ by hand or through a symbolic manipulation language, such as Mathematica or Maple, the fractional calculus provides $\mathbb{E}(X^\alpha)$ ([5]).

$$\mathbb{E}(X^\alpha) = \frac{1}{\Gamma(n-\alpha)} \int_{-\infty}^{0} (-z)^{n-\alpha-1} \frac{d^n}{dz^n} M(z)\, dz,\ n \in \mathbb{N}, \alpha < n.$$

where only real values of $M(t)$ only are needed. Invoking *ME* procedure with constraints $\{\mathbb{E}(X^{\alpha_j})\}_{j=1}^{M}$, similar results as in previous section are obtained.

The knowledge of $M(t)$, $t \in \mathbb{C}$ allows us to calculate $\{\mu_j\}_{j=1}^{M}$ and then $\mathbb{E}(X^\alpha)$ by (7), through an efficient procedure, as suggested by Choudhury ([3]).

As consequence, fractional moments $\mathbb{E}(X^\alpha)$, $\alpha > 0$, may be efficiently calculated through (7).

REFERENCES

1. N.I. Akhieser, The classical moment problem and some related questions in analysis. Hafner, New York (1965).
2. H. Chernoff, A measure of asymptotic efficiency for tests of hypothesis based on the sum of observations. Annals of Mathematical Statistics, **23**, 493-507 (1952).
3. G.L. Choudhury, D.M. Lucantoni, Numerical computation of the moments of a probability distribution from its transform. Operations Research, **44**, n.2, 368-381 (1996).
4. T.M. Cover, J.A. Thomas, Elements of Information Theory. John Wiley & Sons, Inc., New York (1991).
5. N. Cressie, M. Borkent, The moment generating function has its moments. Journal of Statistical Planning and Inference, **13**, 337-344 (1986).
6. D.W. Kammler, Prony's method for completely monotonic functions. J. of Math. Analysis and Applications, **57**, 560-570 (1977).
7. H.K. Kesavan, J.N. Kapur, Entropy Optimization Principles with Applications. Academic Press, New York (1992).
8. B. Klar, On a test for exponentiality against Laplace order dominance. Statistics, **37**, n.6, 505-515 (2003).
9. S. Kullback, A lower bound for discrimination information in terms of variation. IEEE Transaction on Information Theory, **IT-13**, 126-127 (1967).
10. G.D. Lin, Characterizations of Distributions via moments. Sankhya: The Indian Journal of Statistics, **54**, Series A, 128-132 (1992).
11. B.G. Lindsay, P. Basak, Moments determine the tail of a distribution (but not much else). The American Statistician, **54**, n.4, 248-251 (2000).
12. E.P. Merkers, M. Wetzel, A geometric characterization of indeterminate moment sequences. Pacific Journal of Mathematics, **65**, n. 2, 409-419 (1976).
13. P. Naveau, Comparison between the Chernoff and factorial moment bounds for discrete random variables. The American Statistician, **51**, n.1, 40-41 (1997).
14. P.L. Novi Inverardi, A. Tagliani, Maximum entropy density estimation from fractional moments. Communications in Statistics - Theory and Methods, **32**, n.2, 15-32 (2003).
15. T.K. Philips, R. Nelson, The Moment Bound is Tighter That Chernoff's Bound for Positive Tail Probabilities. The American Statistician, **49**, n.2, 175-178 (1995).
16. A. Tagliani, Inverse Z transform and moment problem. Probability in the Engineering and Informational Sciences, **14**, 393-404 (2000).