On the Convergence of the Empirical Mass Function

Ralph P. Russo *, Nariankadu D. Shyamalkumar
Department of Statistics and Actuarial Science
The University of Iowa

Abstract

We show that the empirical mass function associated with a sequence of i.i.d. discrete random variables converges in l^r at the $(n/log_2 n)^{1/2}$ rate, for all $r \geq 2$. For $r < 2$ the rate is shown to fail for heavy tailed distributions. The threshold case of $r = 2$ is explored in detail.

Key words: Law of the iterated logarithm, Empirical probabilities, l^2 distance
1991 MSC: 60F15, 60B12, 62G30

1. Introduction

Let $X, X_1, X_2, \ldots$ be an i.i.d. sequence of discrete random variables taking values in $\mathcal{V} := \{v_1, v_2, \ldots\}$. For $k, n \geq 1$ define $p_k := \Pr(X = v_k)$, and the sample proportion

$$p_{k:n} := \left(\frac{1}{n}\right) \sum_{j=1}^{n} I(X_j = v_k)$$

By Scheffe's theorem, the empirical mass function $\widetilde{p}_n := (p_{1:n}, p_{2:n}, \ldots)$ converges to its limit $\widetilde{p} := (p_1, p_2, \ldots)$ with probability one, under any of the l^r metrics, $1 \leq r \leq \infty$. Of interest is the rate at which this convergence occurs.

For $r > 0$, let $\|\widetilde{x} - \widetilde{y}\|_r$ denote the l^r distance between the vectors $\widetilde{x}$ and $\widetilde{y}$, for $t > 0$ let $\log_2(t)$ denote the modified iterated logarithm $\log\log\max(e^e, t)$, and let $\mathcal{S}_\infty$ denote the ∞-dimensional simplex. If $\widetilde{p}$ has finite support, then the result

$$\limsup_{n\to\infty} \sqrt{\frac{n}{\log_2 n}} \, \|\widetilde{p}_n - \widetilde{p}\|_\infty = \max_k \sqrt{2p_k(1-p_k)} < \infty, \quad a.s. \tag{1}$$

* Corresponding author
Email addresses: rrusso@stat.uiowa.edu (Ralph P. Russo), nshyamal@stat.uiowa.edu (Nariankadu D. Shyamalkumar).

is an easy consequence of the law of the iterated logarithm (LIL) for Bernoulli sequences. That the above holds in general follows from Olshen and Siegmund (1971), suggesting that

$$L\left(r;\widetilde{p}\right) := \limsup_{n\to\infty} \sqrt{\frac{n}{\log_2 n}} \left\|\widetilde{p}_n - \widetilde{p}\right\|_r \tag{2}$$

is a quantity of interest. We observe that $L\left(r;\widetilde{p}\right)$ is almost surely a constant (possibly infinite) by the *Hewitt-Savage zero-one law*. For convenience, we use $L\left(r;\widetilde{p}\right)$ to represent this constant. Our interest here is primarily in $L\left(2;\widetilde{p}\right)$, the reason being that $r = 2$ is the threshold value of r at which $\sup\{L\left(r;\widetilde{p}\right)|\ \widetilde{p} \in \mathcal{S}_\infty\}$ is finite. Specifically, we show in section 2 (see also statement (10)) that

$$\sup_{r\geq 2;\widetilde{p}} L\left(r;\widetilde{p}\right) \leq 2^{(4-r)/2r} \tag{3}$$

while, for suitably heavy tailed distributions, $L\left(r;\widetilde{p}\right) = \infty$ for all $r < 2$. In section 3 we show that $L\left(2;\widetilde{p}\right)$ is uniquely maximized over $\mathcal{S}_\infty$ (rather surprisingly) at the uniform distribution on two points. In section 4 we provide an error bound for a natural approximation of $L\left(2;\widetilde{p}\right)$, to any desired level of accuracy.

For convenience we will assume (with no loss) that p_k is non-increasing in k. By $<\cdot,\cdot>$ we denote the inner product of l^2. We use $\|\cdot\|$ for the operator norm (this should be clear from the context). If $\tilde{x}$ is a vector with i-th component x_i, by $\sqrt{\tilde{x}}$ (resp., $\tilde{x}^\alpha$) we denote the vector whose i-th component is $\sqrt{x_i}$ (resp., x_i^α).

2. The Finiteness of $L\left(2;\widetilde{p}\right)$ on $\mathcal{S}_\infty$

Our first result establishes the $(n/log_2 n)^{1/2}$ rate of l^2 convergence of the empirical mass function to its limit, and provides an exact expression for $L\left(2;\widetilde{p}\right)$. This exact expression, while not in closed form, can nonetheless be bounded by a finite constant, independent of $\widetilde{p}$ in S_∞. This, together with statement (1), yields a uniform bound on $L(r;\widetilde{p})$ over $\widetilde{p}$ in S_∞ and $r \geq 2$. On the other hand, Example 1 will show that for heavy tailed $\widetilde{p}$, $L(r;\widetilde{p}) = \infty$ for all $r < 2$.

Theorem 1. *We have for $\widetilde{p}$ in S_∞,*

$$L\left(2;\widetilde{p}\right) = \left(2\sup\left\{\sum_{k\geq 1} p_k^2 \left(a_k - \sum_{l\geq 1} p_l a_l\right)^2 \Big| \sum_{k\geq 1} p_k a_k^2 \leq 1\right\}\right)^{\frac{1}{2}} < \infty \tag{4}$$

Moreover, we have statement (3).

PROOF. Define the kernel $h(\cdot,\cdot)$ by

$$h(i,j) = I_{\{i=j\}} - (p_i + p_j) + \sum_{k\geq 1} p_k^2, \quad i,j = 1,2,\ldots$$

We have

$$\left(\frac{n}{\log_2 n}\right)\sum_{j=1}^{\infty}(p_j^n - p_j)^2 = \left(\frac{2}{n\log_2 n}\right)\sum_{1\le i<j\le n} h(X_i, X_j) + \left(\frac{1}{\log_2 n}\right)\left(1+\sum_{k\ge 1} p_k^2\right) - \left(\frac{2}{\log_2 n}\right)\left(\frac{\sum_{i=1}^n p_{X_i}}{n}\right) = A_n + B_n + C_n$$

It is easy to check that A_n is a canonical U-statistic of order 2. By (Dehling, 1989) we have

$$\limsup A_n = 2\sup\left\{\sum_{k\ge 1} p_k^2\left(a_k - \sum_{l\ge 1} p_l a_l\right)^2 \Big| \sum_{k\ge 1} p_k a_k^2 \le 1\right\} \ a.s.$$

As $n\to\infty$, B_n vanishes trivially, while C_n vanishes in the almost sure sense by the strong law of large numbers. Hence we have (4). Towards showing (3) we observe that by (1) we have,

$$\begin{aligned}\limsup \sqrt{\frac{n}{\log_2 n}}\left\|\widetilde{p}_n - \widetilde{p}\right\|_r &\le L\left(2;\widetilde{p}\right)^{2/r}\limsup\left(\sqrt{\frac{n}{\log_2 n}}\left\|\widetilde{p}_n-\widetilde{p}\right\|_\infty\right)^{\frac{(r-2)}{r}} \\ &= L\left(2;\widetilde{p}\right)^{2/r}\left(2\max_k(p_k(1-p_k))\right)^{\frac{(r-2)}{2r}}\end{aligned} \tag{5}$$

Also, as a consequence of $\mathrm{Var}\,(Y)\le \mathbb{E}\big(Y^2\big)$ we have

$$\sum_{k\ge 1} p_k^2\left(a_k - \sum_{l\ge 1} p_l a_l\right)^2 \le \sum_{k\ge 1} p_k a_k^2. \tag{6}$$

From (6) we have $L\left(2;\widetilde{p}\right)\le\sqrt{2}$, and this with (5) yields (3). □

Remark 1. The proof in Olshen and Siegmund (1971) establishes also the $(n/log_2 n)^{1/2}$ rate under l^r for $r\ge 4$. In the case when $\widetilde{p}$ has a finite support, the problem can be tackled for all l^r $(r>0)$ using the LIL for random vectors taking values in $\mathbb{R}^k$ (in a manner similar to the proof of Lemma 3 in Finkelstein (1971)).

Example 1. Suppose that the tail of $\widetilde{p}$ is sufficiently heavy so that

$$\left(\frac{n}{\log_2 n}\right)^{\frac{r}{2}}\sum_{k>n}(p_k)^r \to\infty, \quad \forall r<2. \tag{7}$$

Then with $\mathcal{H}_n = \mathcal{V} - \{X_1,...,X_n\}$, we have

$$\sqrt{\frac{n}{\log_2 n}}\left\|\widetilde{p}_n-\widetilde{p}\right\|_r \ge \sqrt{\frac{n}{\log_2 n}}\left(\sum_{k\in\mathcal{H}_n}(p_k)^r\right)^{\frac{1}{r}} \ge \sqrt{\frac{n}{\log_2 n}}\left(\sum_{k>n}(p_k)^r\right)^{\frac{1}{r}} \to\infty \tag{8}$$

and thus $L\left(2;\widetilde{p}\right)=\infty$ for all $r<2$. For an example of a suitably heavy tailed $\widetilde{p}$, take $p_k\propto\left(k[\log k]^2\right)^{-1}$ for $k\ge 1$. □

3. The Maximizer of $L(2;\widetilde{p})$ on $\mathcal{S}_\infty$

Towards finding the maximizer of $L(2;\widetilde{p})$ in $\mathcal{S}_\infty$ we define the function $\psi(\cdot)$ on the space S_∞ of mass functions as

$$\psi(\tilde{p}) := \sup\left\{ \sum_{k\geq 1} p_k^2 \left(a_k - \sum_{l\geq 1} p_l a_l \right)^2 \Big| \sum_{k\geq 1} p_k a_k^2 \leq 1 \right\}$$

and we observe that $L(2;\widetilde{p}) = \sqrt{2\psi(\widetilde{p})}$.

By $S_{\tilde{p}}$ we denote the operator from l^2 to l^2 defined by

$$S_{\tilde{p}}(\tilde{x}) = \left\{ p_i x_i + \left[\sqrt{p_i} \sum_{j\geq 1} p_j^2 - p_i^{3/2} \right] < \sqrt{\tilde{p}}, \tilde{x} > - \sqrt{p_i} < \tilde{p}^{3/2}, \tilde{x} > \right\}_{i\geq 1},$$

for $\tilde{x} \in l^2$. Our interest in $S_{\tilde{p}}$ arises from $\psi(\tilde{p}) = \sup_{\|\tilde{x}\|=1} \tilde{x}' S_{\tilde{p}}(\tilde{x})$. It is easy to check that $S_{\tilde{p}}$ is a Hilbert-Schmidt operator and that its Hilbert-Schmidt squared norm (see section XI.6 of (Dunford and Schwartz, 1963)) is given by

$$\sum_{j\geq 1} p_j^2(1-2p_j) + \left(\sum_{j\geq 1} p_j^2 \right)^2.$$

As every Hilbert-Schmidt operator is compact (Theorem 6 of section XI.6 in (Dunford and Schwartz, 1963)) and $S_{\tilde{p}}$ is self adjoint, we have that $S_{\tilde{p}}$ is a compact self adjoint operator. Now by Theorem 3 of section X in (Dunford and Schwartz, 1963) we have,

$$\psi(\tilde{p}) = \sup_{\|\tilde{x}\|=1} \tilde{x}' S_{\tilde{p}}(\tilde{x}) = \lambda,$$

where λ is the maximum eigenvalue of $S_{\tilde{p}}$. We note also that $\sqrt{\tilde{p}}$ is a zero eigenvector of $S_{\tilde{p}}$, *i.e.* $S_{\tilde{p}}(\sqrt{\tilde{p}}) = 0$.

Before delving further, it is instructive to calculate $L(2;\widetilde{p})$ for some known distributions with finite support. In these cases, $S_{\tilde{p}}$ is a symmetric non-negative definite matrix.

Example 2 (Two Point Support). The simplest of discrete distributions are the ones with a two point support. In this case by the usual law of the iterated logarithm we have $L_p = 4pq$ where p and q are the probabilities assigned to the two points ($q := 1-p$). The result also follows by observing that

$$S_{\tilde{p}} = 2pq \begin{pmatrix} q & -\sqrt{pq} \\ -\sqrt{pq} & p \end{pmatrix},$$

with trace of $2pq$ and a single zero eigenvalue, implying $L(2;\widetilde{p}) = \sqrt{2\psi(\widetilde{p})} = 2\sqrt{pq}$. □

Example 3 (Discrete Uniform Distribution). Consider a uniform distribution on $m(\geq 1)$ points, of interest as it maximizes the trace of $S_{\tilde{p}}$ among all m-point support distributions. In this case the matrix $S_{\tilde{p}}$ can be written succinctly as

$$S_{\tilde{p}} = \left(\frac{1}{m}\right) I_{m\times m} - \left(\frac{1}{m^2}\right) \mathbf{1_{m\times 1}} \mathbf{1}'_{\mathbf{m\times 1}}$$

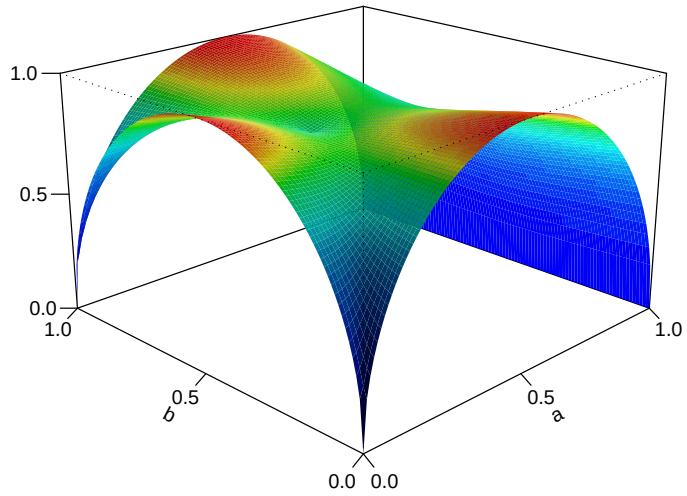


Fig. 1: $L\left(2;\widetilde{p}\right)$: three point support

The above makes it is clear that the vectors in the ($m-1$ dimensional) space orthogonal to $1_{m\times 1}$ are eigenvectors with eigenvalue $1/m$, implying a maximum eigenvalue of $1/m$ and $L\left(2;\widetilde{p}\right) = \sqrt{2/m}$. This result also follows from Lemma 3 of (Finkelstein, 1971). Interestingly, among all m-point support distributions that have $tr(S_{\tilde{p}}) = 1-(1/m)$, the uniform on m points minimizes the maximum eigenvalue. □

Example 4 (Three Point Support). Consider an arbitrary discrete distribution supported on three points. In this case it can be checked that the two non-zero eigenvalues of the matrix $S_{\tilde{p}}$ are

$$\left(\frac{1}{2}\right)\left[tr(S_{\tilde{p}}) \mp \sqrt{2\sum_{i=1}^{3} p_i^4 - \sum_{i=1}^{3} p_i^2 + 6\sum_{1\le i\neq j\le 3} p_i p_j^2 - 2p_1p_2p_3 - 1}\right]$$

Figure 1 is an image plot of $L\left(2;\widetilde{p}\right)$ as a function of a and b of the *stick breaking* parametrization of a three point distribution, *i.e.*

$$p_1 = a,\ p_2 = (1-a)b,\ \text{and}\ p_3 = (1-a)(1-b).$$

Noteworthy is that the maximum value of $L\left(2;\widetilde{p}\right)$ is 1, attained at $(1/2, 0)$, $(1/2, 1)$ and $(0, 1/2)$ - all representing the uniform distribution on two points.

□

Example 4 shows that the maximum value of $L(2;\widetilde{p})$ among 3-point distributions is attained at a 2-point distribution, leading one to conjecture that the maximum of $L(2;\widetilde{p})$ among *all* discrete distributions is attained at the uniform distribution on two points. This conjecture is settled in the affirmative by the following result:

Theorem 2. *For all $\widetilde{p}$ in $\mathcal{S}_\infty$ we have $0 \leq L(2;\widetilde{p}) \leq 1$. If $L(2;\widetilde{p}) = 0$ then $\widetilde{p}$ is a degenerate distribution. If $L(2;\widetilde{p}) = 1$ then $\widetilde{p}$ is the uniform distribution on two points.*

PROOF. Without loss of generality we will assume that at least two components of $\tilde{p}$ are positive. Let $\tilde{z}$ be an eigenvector of $S_{\tilde{p}}$ which corresponds to its maximum eigenvalue, say λ. Since $\sqrt{\tilde{p}}$ is in the null space of $S_{\tilde{p}}$ and $S_{\tilde{p}}$ is self-adjoint, we have $< \sqrt{\tilde{p}}, \tilde{z} >= 0$ (see for example, Theorem 9.1-1 of (Kreyszig, 1978)). This implies,

$$z_i(p_i - \lambda) = \sqrt{p_i} < z, \tilde{p}^{3/2} >, \quad i = 1, 2, \ldots \tag{9}$$

and moreover that $\lambda = (\sum_{i\geq 1} z_i^2 p_i)/\sum_{i\geq 1} z_i^2$. From the last expression it follows easily that $\lambda \leq \max_{i\geq 1} p_i$. Now we show that if $\lambda = p_j$ for some $j \geq 1$ then $\lambda \leq (1/2)$. This is so as $\lambda = p_j$ for some $j \geq 1$ and (9) imply that $< z, \tilde{p}^{3/2} >= 0$. Hence $z_i = 0$ for all i such that $p_i \neq p_j$. This along with $< \sqrt{\tilde{p}}, \tilde{z} >= 0$ and $\tilde{z} \neq 0$ implies that the set $\{i \geq 1 | p_i = p_j\}$ has at least two points. Hence $2p_j \leq 1$ or $\lambda \leq (1/2)$.

The only case we need to rule out is that of $p_1 > \lambda > (1/2) > p_j$ for $j \geq 2$. Suppose this case holds. Defining $\alpha := z_1^2/\sum_{i\geq 1} z_i^2$, we have $(1/2) < \lambda \leq \alpha p_1 + (1-\alpha)p_2$, which implies $(1-\alpha) < [p_1 - (1/2)]/(p_1 - p_2)$. As $p_1 + p_2 \leq 1$ and $p_1 > p_2$ we have $[p_1 - (1/2)]/(p_1 - p_2) \leq 1/2$, and hence that $\alpha > 1/2$. Finally, by using the Cauchy-Schwartz inequality and the fact that $< \sqrt{\tilde{p}}, \tilde{z} >= 0$ implies $z_1\sqrt{p_1} = -\sum_{j\geq 2} \sqrt{p_j} z_j$, we have

$$\lambda \leq \frac{p_1 z_1^2 + p_2 \sum_{j\geq 2} z_j^2}{\sum_{i\geq 1} z_i^2} \leq \frac{\sum_{j\geq 2} p_j \sum_{j\geq 2} z_j^2 + p_2 \sum_{j\geq 2} z_j^2}{\sum_{i\geq 1} z_i^2}$$

$$= (1 - (p_1 - p_2))(1 - \alpha) < \frac{1}{2}$$

A contradiction. Hence the proof. □

Remark 2. We note that the upper bound of 1 on $L(2;\widetilde{p})$, provided by the above theorem, leads to an improvement of statement (3):

$$\sup_{r\geq 2;\widetilde{p}} L(r;\widetilde{p}) \leq 2^{(2-r)/2r} \tag{10}$$

4. An Approximation of $L(2;\widetilde{p})$

Consider the problem of estimation of the mass function using a random sample of size n from a distribution with support contained in the set of integers. Since quite generally the empirical mass function is asymptotically optimal in the sense of mean integrated squared error, see (Watson and Leadbetter, 1963), it is an estimator of interest. Hence the limit superior in (2) with $r = 2$, viewed as an asymptotic *worst case* measure of estimation error incurred by the empirical mass function, is a quantity of interest. This is one motivation to compute the value of $L(2;\widetilde{p})$.

Example 4 suggests that a closed form solution for $L(2;\widetilde{p})$ is unattainable. For a given distribution with finite support, one can resort to numerically computing $L(2;\widetilde{p})$. For distributions with infinite support, an approximation is needed. The following theorem gives one such approximation along with an error bound. First we need to define the operator $T_{\tilde{p}}$ from l^2 to l^2:

$$T_{\tilde{p}}(\tilde{x}) := \left\{\sqrt{p_i}x_i - p_i < \sqrt{\tilde{p}}, \tilde{x} >\right\}_{i\geq 1}, \quad \tilde{x} \in l^2$$

Note that $T_{\tilde{p}}$ is related to $S_{\tilde{p}}$ by $S_{\tilde{p}} = T'_{\tilde{p}} T_{\tilde{p}}$. It follows then that $\psi(\tilde{p})$ is the squared norm of the operator $T_{\tilde{p}}$. Now consider the canonical basis for l^2 and let P_n be the projection onto the span of its first n elements. It is easily seen that $T_{\tilde{p}:n}$ defined as $P_n T_{\tilde{p}} P_n$ is such that if $\tilde{y} = T_{\tilde{p}:n}(\tilde{x})$ then $\tilde{y}$ is given by

$$y_i = \begin{cases} \sqrt{p_i}x_i - p_i \sum_{j=1}^{n} \sqrt{p_j}x_j & i = 1, \ldots, n; \\ 0 & i > n; \end{cases}.$$

Finally, let $S_{\tilde{p}:n}$ be defined as $T'_{\tilde{p}:n} T_{\tilde{p}:n}$.

Theorem 3. *For any probability distribution $\tilde{p}$ we have*

$$L(2;\widetilde{p}) = \lim_{n\to\infty} \sqrt{2 \sup_{\|\tilde{x}\|=1} \tilde{x}' S_{\tilde{p}:n}(\tilde{x})} \tag{11}$$

Moreover,

$$\left| L(2;\widetilde{p}) - \sqrt{2 max.\ eigenvalue\ of\ S_{\tilde{p}:n}} \right| \leq \sqrt{10 \sum_{j>n} p_j}, \quad n = 1, 2, \ldots \tag{12}$$

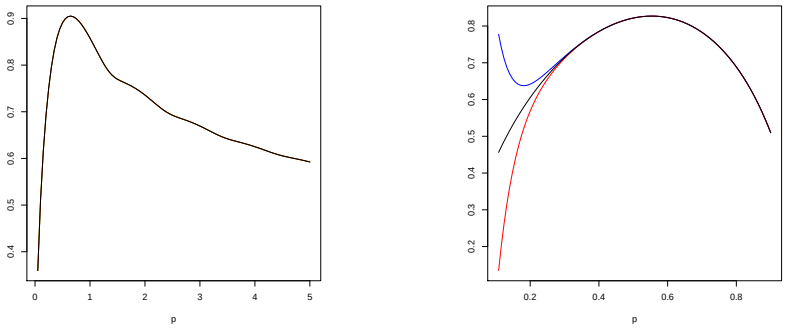
PROOF. Let $(T_{\tilde{p}} - T_{\tilde{p}:n})(\tilde{x}) = \tilde{y}$, where $\tilde{y}$ is given by

$$y_i = \begin{cases} -p_i \sum_{j>n} \sqrt{p_j}x_j, & i = 1, \ldots, n; \\ \sqrt{p_i}x_i - p_i < \sqrt{\tilde{p}}, \tilde{x} >, & i > n; \end{cases}.$$

Now with multiple use of the Cauchy-Schwartz inequality we have,

$$\begin{aligned} \|\tilde{y}\|^2 &= \left(\sum_{j=1}^{n} p_j^2\right) \left[\sum_{j>n} \sqrt{p_j}x_j\right]^2 \\ &\quad + \sum_{j>n} p_j x_j^2 + \left(< \sqrt{\tilde{p}}, \tilde{x} >\right)^2 \sum_{j>n} p_j^2 - 2 < \sqrt{\tilde{p}}, \tilde{x} > \sum_{j>n} p_j^{3/2} x_j \\ &\leq 5 \left(\sum_{j>n} p_j\right) \|\tilde{x}\|^2 \end{aligned}$$

The above implies that $\|(T_{\tilde{p}} - T_{\tilde{p}:n})(\tilde{x})\|^2 / \|\tilde{x}\|^2$ is bounded above by $5 \sum_{j>n} p_j$; in other words $\|T_{\tilde{p}} - T_{\tilde{p}:n}\| \leq \sqrt{5 \sum_{j>n} p_j}$. This implies (12) which in turn implies (11). Hence the proof. □



(a) Poisson Distribution (b) Geometric Distrbution

Fig. 2: The limit superior as a function of the parameter

Example 5 (Poisson and Geometric Distributions). As an example of a discrete distribution with infinite support we consider the Poisson and geometric distributions. For these the approximation to $L\left(2;\widetilde{p}\right)$ along with accompanying error bounds derived from Theorem 3 are plotted in Figure 2. The approximation was based on the first 40 (i.e. with $n = 40$ in Theorem 3) terms of these distributions; the error bounds become wider as the mass moves to the right, corresponding to low p values in the case of the geometric and high λ values in the case of the Poisson. We note that this can be improved upon by considering the n most probable values rather than the initial n values. The latter uses the fact that $L\left(2;\widetilde{p}\right)$ is invariant to any one-to-one transformation of the distribution.

Note that the maximum of $L\left(2;\widetilde{p}\right)$ is attained in both cases in the proximity of the parameter value corresponding to $p_1 = 0.5$ (corresponding to the probability at 0). In this sense, the plot in view of Theorem 2 is not very surprising, even though not entirely obvious either. □

References

Dehling, H., 1989. Complete convergence of triangular arrays and the law of the iterated logarithm for U-statistics. Statist. Probab. Lett. 7 (4), 319–321.

Dunford, N., Schwartz, J. T., 1963. Linear Operators, Spectral Theory, Self Adjoint Operators in Hilbert Space, Part 2. Wiley-Interscience.

Finkelstein, H., 1971. The law of the iterated logarithm for empirical distributions. Ann. Math. Statist. 42, 607–615.

Kreyszig, E., 1978. Introductory Functional Analysis with Applications. Wiley.

Olshen, R., Siegmund, D., 1971. On the maximum likelihood estimate of cell probabilities. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 19, 52–56.

Watson, G. S., Leadbetter, M. R., 1963. On the estimation of the probability density. I. Ann. Math. Statist. 34, 480–491.