

# A new empirical process with applications <sup>(\*)</sup>

## *Un nuovo processo empirico con applicazioni*

Mauro Gasparini and Gianfranco Genta

Dipartimento di Matematica, Politecnico di Torino

e-mail: gasparini@calvino.polito.it

**Riassunto:** Data l'unione dell'insieme  $\{0, 1\}$  e di un insieme di  $n$  punti nell'intervallo  $(0, 1)$ , si definisca la proiezione di un punto  $x \in [0, 1]$  come l'elemento, appartenente alla suddetta unione, più vicino a  $x$ . Se gli  $n$  punti sono realizzazioni di una successione di variabili aleatorie i.i.d. uniformi, per  $x \in [0, 1]$  rimane allora definito il processo proiezione empirico. Di tale processo vengono dimostrate alcune proprietà e ne viene fatto il confronto con il processo distributivo empirico e il processo quantile empirico. La motivazione per la definizione del nuovo processo empirico sorge in un problema di ricerca operativa quando si voglia studiare la distribuzione della distanza minima che un concorrente gettato a caso su una strada, disseminata di premi equivalenti in posizioni casuali, deve percorrere per raggiungere il premio più vicino.

**Keywords:** empirical process, Voronoi tessellation, spacing.

### 1. Definition of the empirical projection process

Let  $A_1, A_2, \dots$ , be a sequence of i.i.d. random variables from a uniform distribution on the interval  $[0, 1]$  and, for each integer  $n > 0$ , let  $A_{(1)}^n \leq \dots \leq A_{(n)}^n$  be their order statistics. Furthermore, let  $A_{(0)}^n = 0$  and  $A_{(n+1)}^n = 1$  for a convenient notation. For given  $n$  and  $x \in [0, 1]$ , let  $G_n(x)$  be the element of the finite set  $\{0, 1, A_1, \dots, A_n\}$  closest to  $x$ , that is the value such that

$$|G_n(x) - x| = \min_{1 \leq i \leq n} |A_i - x|. \quad (1)$$

$G_n(x)$  is the projection of  $x$  on the finite set  $\{0, 1, A_1, \dots, A_n\}$ . By simple geometrical considerations, it is easy to see that

$$G_n(x) = \begin{cases} 0 & \text{if } 0 \leq x < \frac{A_{(1)}^n}{2} \\ \vdots & \\ A_{(j)}^n & \text{if } \frac{A_{(j-1)}^n + A_{(j)}^n}{2} \leq x < \frac{A_{(j)}^n + A_{(j+1)}^n}{2} \\ \vdots & \\ 1 & \text{if } \frac{A_{(n)}^n + 1}{2} \leq x \leq 1 \end{cases} \quad (2)$$

that is,  $G_n(x)$  is a step function taking on different values on the one-dimensional Voronoi tessellation of  $[0, 1]$  induced by  $A_1, A_2, \dots, A_n$ . Let  $(E)$  denote the indicator function of an event  $E$  and define the *spacings*  $D_j^n = A_{(j)}^n - A_{(j-1)}^n, j = 1, 2, \dots, n+1$ , reviewed

---

<sup>(\*)</sup> Work supported by MIUR, grant PRIN2005 "Statistical tools in system biology". The authors wish to thank Giovanni Pistone for useful conversations.

for example by Pyke (1965). Rewriting expression (2) more compactly, then summing by parts, it is possible to obtain several alternative representations:

$$\begin{aligned}
G_n(x) &= \sum_{j=1}^n \left( \frac{A_{(j-1)}^n + A_{(j)}^n}{2} \leq x < \frac{A_{(j)}^n + A_{(j+1)}^n}{2} \right) \cdot A_{(j)}^n + \\
&\quad \left( \frac{A_{(n)}^n + 1}{2} \leq x \leq 1 \right) \\
&= \sum_{i=0}^n \left( \frac{A_{(i)}^n + A_{(i+1)}^n}{2} \leq x \right) \cdot [A_{(i+1)}^n - A_{(i)}^n] \\
&= \sum_{i=0}^n \left( A_{(i)}^n + \frac{D_{i+1}^n}{2} \leq x \right) \cdot D_{i+1}^n. \tag{3}
\end{aligned}$$

For varying  $x \in [0, 1]$ ,  $G_n(\cdot)$  defines a new stochastic process which is the object of this note and can be called *(uniform) empirical projection process*.

## 2. Properties of the empirical projection process

The projection process enjoys some degree of distributional symmetry.

**Theorem 1** *For each  $x \in [0, 1]$ , the random variable  $1 - G_n(1 - x)$  has the same distribution as  $G_n(x)$ . In symbols:  $1 - G_n(1 - x) \stackrel{d}{=} G_n(x)$ , where  $\stackrel{d}{=}$  denotes equality of the distributions.*

**Proof**

Let  $B_j = 1 - A_j$ ,  $j = 1, 2, \dots$  and notice that, since each  $A_j$  is uniform,  $B_j \stackrel{d}{=} 1 - A_j$ ,  $j = 1, 2, \dots$ . For each integer  $n > 0$  define the order statistics  $0 = B_{(0)}^n \leq B_{(1)}^n \leq \dots \leq B_{(n)}^n \leq B_{(n+1)}^n = 1$  and let  $G_n^B(x)$ ,  $x \in [0, 1]$  be the relative empirical projection process. Since  $B_{(j)}^n = 1 - A_{(n-j+1)}^n$ ,  $j = 0, 1, \dots, n+1$ , we can write

$$\begin{aligned}
G_n^B(x) &= \sum_{k=0}^n \left( \frac{B_{(k)}^n + B_{(k+1)}^n}{2} \leq x \right) \cdot [B_{(k+1)}^n - B_{(k)}^n] \\
&= \sum_{k=0}^n \left( \frac{A_{(n-k+1)}^n + A_{(n-k)}^n}{2} \geq 1 - x \right) \cdot [A_{(n-k+1)}^n - A_{(n-k)}^n] \\
&= \sum_{k=0}^n \left[ 1 - \left( \frac{A_{(n-k+1)}^n + A_{(n-k)}^n}{2} < 1 - x \right) \right] \cdot [A_{(n-k+1)}^n - A_{(n-k)}^n] \\
&= 1 - \sum_{k=0}^n \left( \frac{A_{(n-k+1)}^n + A_{(n-k)}^n}{2} < 1 - x \right) \cdot [A_{(n-k+1)}^n - A_{(n-k)}^n] \\
&= 1 - \sum_{i=0}^n \left( \frac{A_{(i)}^n + A_{(i+1)}^n}{2} < 1 - x \right) \cdot [A_{(i+1)}^n - A_{(i)}^n] \\
&= 1 - G_n(1 - x) \quad \text{a.s.}
\end{aligned}$$

But  $G_n^B(x) \stackrel{d}{=} G_n(x)$ , being both uniform empirical projections, henceforth the statement.

The exact distribution of  $G_n(x)$  is not simple, but expression (3) allows one to calculate its expected value, as in the following theorem.

**Theorem 2** For each  $x \in [0, 1]$ , the expected value of the random variable  $G_n(x)$  is the following polynomial of degree  $(n + 1)$ :

$$\mathbb{E}(G_n(x)) = \begin{cases} x - x(1 - 2x)^n & \text{if } 0 \leq x \leq \frac{1}{2} \\ x + (1 - x)(1 - 2(1 - x))^n & \text{if } \frac{1}{2} < x \leq 1 \end{cases} \quad (4)$$

**Proof**

By Theorem 1, it is enough to consider the case  $0 \leq x \leq 1/2$ . For such an  $x$ ,

$$\begin{aligned} \mathbb{E}(G_n(x)) &= \mathbb{E} \left( \left( A_{(0)}^n + \frac{D_1^n}{2} \leq x \right) \cdot D_1^n \right) + \\ &\sum_{i=1}^{n-1} \mathbb{E} \left( \left( A_{(i)}^n + \frac{D_{i+1}^n}{2} \leq x \right) \cdot D_{i+1}^n \right) + \mathbb{E} \left( \left( A_{(n)}^n + \frac{D_{n+1}^n}{2} \leq x \right) \cdot D_{n+1}^n \right) \end{aligned}$$

The three terms on the right hand side can be evaluated separately. For the first one,

$$\begin{aligned} \mathbb{E} \left\{ \left( A_{(0)}^n + \frac{D_1^n}{2} \leq x \right) \cdot D_1^n \right\} &= \mathbb{E} \{ (A_{(1)}^n \leq 2x) \cdot A_{(1)}^n \} \\ &= \int_0^{2x} n(1-t)^{n-1} t dt = \frac{1 - (1-2x)^n(1+2nx)}{n+1} \end{aligned} \quad (5)$$

since it is known that  $A_{(1)}^n \sim \text{Beta}(1, n)$ . For the second one, it is possible to see that the joint density of  $A_{(i)}^n$  and  $D_{i+1}^n$  is

$$\begin{aligned} f_{(A_{(i)}^n, D_{i+1}^n)}(s, t) &= \frac{n!}{(i-1)!(n-i-1)!} s^{i-1} (1 - (s+t))^{n-i-1} \\ &\quad (0 < s < 1)(0 < t < 1-s), \end{aligned}$$

based on which we obtain

$$\begin{aligned} \sum_{i=1}^{n-1} \mathbb{E} \left( \left( A_{(i)}^n + \frac{D_{i+1}^n}{2} \leq x \right) \cdot D_{i+1}^n \right) &= \\ \sum_{i=1}^{n-1} \int_0^{2x} \int_0^{(x-\frac{t}{2})} \frac{n!}{(i-1)!(n-i-1)!} s^{i-1} (1 - (s+t))^{n-i-1} t ds dt &= \\ \int_0^{2x} \int_0^{(x-\frac{t}{2})} n(n-1)t(1-t)^{n-2} ds dt &= \\ \frac{-1 + (n+1)x + (1-2x)^n[(n-1)x+1]}{n+1} \end{aligned} \quad (6)$$

by Newton's binomial formula. The third term is 0. Summing expressions (5) and (6), we obtain expression (4).

Notice that

$$\lim_{n \rightarrow +\infty} \mathbb{E} \{ G_n(x) \} = x,$$

which shows that  $G_n(x)$  is an asymptotically unbiased estimator of the underlying distribution function, which in the uniform case coincides with the quantile function. Other asymptotic properties of the empirical projection process and its consequences for statistical estimation and testing will be studied elsewhere.

### 3. Comparisons to other empirical processes

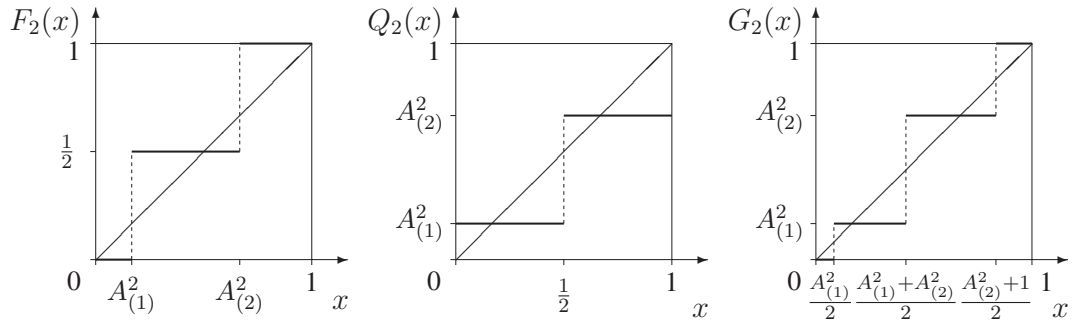
The process  $G_n(\cdot)$  is a step function with random jumps equal to the spacings  $D_{j+1}^n$  at random points equal to the order statistic midpoints  $(A_{(j)}^n + A_{(j+1)}^n)/2$ , for  $j = 0, 1, \dots, n$ . It differs therefore substantially from the (uniform) *empirical distribution function process*, defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n (A_i \leq x), \quad x \in [0, 1]$$

but also from the (uniform) *empirical quantile process*, defined usually as

$$Q_n(x) = \inf\{y : F_n(y) \geq x\}, \quad x \in [0, 1]$$

(notice that for the uniform sampling it is possible to define the quantile function also for the boundary points 0 and 1). With probability 1, the process  $F_n(\cdot)$  has fixed jumps of sizes  $1/n$  at random points equal to the order statistics  $A_{(j)}^n$ , whereas the process  $Q_n(\cdot)$  has random jumps of sizes equal to the spacings  $D_{j+1}^n = A_{(j+1)}^n - A_{(j)}^n$  at fixed points  $j/n$ , for  $j = 0, 1, \dots, n$ .



**Figure 1:** Three uniform empirical processes given  $A_{(1)}^n = 1/6$  and  $A_{(2)}^n = 2/3$ : from left to right, the distribution process, the quantile process and the projection process.

### 4. Motivations for the empirical projection process

Applications of the projection process arise in a variety of contexts, as the Voronoi tessellations do. The simplest representation involves the following problem in Operation Research. Suppose a series of equivalent prizes are located at random on a linear path of unit size and suppose a player is plunged at a random location on the path. Then the distance the player has to run in order to reach the closest prize is  $|X - G_n(X)|$  where  $G_n(\cdot)$  is the empirical projection process relative to the random locations of the prizes  $A_1, A_2, \dots$ , and  $X$  is a uniform random variable independent of  $A_1, A_2, \dots$ , representing the location of the player.

### References

- Pyke R. (1965) Spacings (with discussion), *Journal of the Royal Statistical Society, Series B, Methodological*, 27, 395-449.