Fast computation of boundary crossing probabilities for the empirical CDF

Amit Moscovich-Eiger, Boaz Nadler, Weizmann Institute of Science, Israel. amit.moscovich@weizmann.ac.il; boaz.nadler@weizmann.ac.il

Problem setup

Let $F_n$ be the empirical CDF of $n$ draws from $U[0,1]$ w.l.o.g. Given two functions $g,h: \mathbb{R} \to \mathbb{R}$, compute the non-crossing probability

$$P_r \{ g(t) < F_n(t) < h(t) \}$$

Several algorithms have been proposed over the years, all are $O(n^2)$.


Equivalent formulation

Let $U_1 \leq U_2 \leq \ldots \leq U_n$ be the order statistics of $n$ draws from $U[0,1]$. Given arbitrary bounds $b_1,\ldots,b_n, d_1,\ldots,d_n \in \mathbb{R}$ compute the probability

$$P_r \{ |g(U_i) - b_i| < |h(U_i) - d_i| \}$$

Two-sided $O(n^3)$ algorithm [F&S 1998]

**Lemma.** $F_n$ satisfies $g(t) < F_n(t) < h(t)$ for all $t$ if only if it satisfies these inequalities at all times when $n/2 < g(t) + h(t)$ cross an integer.

**Definition.** For any $s \in [0,1]$ and any $m \in \{0,1,2,\ldots\}$, let

$$R(s,m) := P_r \{ g(t) < F_n(t) < h(t) \}$$

Recursion relations. Let $0 = t_0 \leq t_1 \leq \ldots \leq t_k = 1$ denote the sorted set of integer-crossing times of $g(t)$ and $h(t)$. The Chapman-Kolmogorov equations give the recursion relations of [1]:

$$R(t_{i+1},m) = \sum_{t_i} R(t_i,m) \cdot P_r \{ |g(t_i) - t_i| = |h(t_i) - t_i| \}$$

where $P_r \{ |g(t_i) - t_i| = |h(t_i) - t_i| \} = \Pr \{ \text{Binomial}(n, t_i - \frac{m}{t_i}, \frac{h(t_i) - t_i}{t_i}) = m \}$.

**Solution.** Eq. (1) is equal to $R(1,n)$, which can be computed in $O(n^2)$.

Two-sided $O(n^3)$ algorithm [K&S 2001]

**Lemma.** The distribution of the stochastic process $n \cdot F_n(t)$ is identical to that of a Poisson process $\xi(t)$ with intensity $n$ conditioned on $\xi(1) = n$.

**Definition.** For any $s \in [0,1]$ and any $m \in \{0,1,2,\ldots\}$, let

$$Q(s,m) := \Pr \{ g(t) < \xi(t) < h(t) \}$$

Recursion relations. Similarly to the previous algorithm, the Chapman-Kolmogorov equations give the recursion relations of [2]:

$$Q(t_{i+1},m) = \sum_{t_i} Q(t_i,m) \cdot P_r \{ |Z_i - m - t_i| = 0 \}$$

where $Z_i$ is a Poisson random variable with intensity $m + t_i - t_i$.

**Solution.** Apply the lemma to obtain

$$P_r \{ g(t) < F_n(t) < h(t) \} = Q(1,n) / \Pr \{ \text{Poisson}(n) = n \}.$$  
Computing $Q(1,n)$ still requires $O(n^3)$ operations.

New two-sided $O(n^3 \log n)$ algorithm

Denote $Q_n = (Q(t_0,0),\ldots,Q(t_n,0),\ldots)$ and $\pi_n = (\Pr \{ Z_i = 0 \}, \ldots , \Pr \{ Z_i = n \})$. $Z_{\lambda}$ is a Poisson random variable with expected value $\lambda$.

**Key idea:** the vector $Q_n$ is nothing but a truncated linear convolution of $Q_1$ and $\pi_n$. Hence, using the circular convolution theorem and the Fast Fourier Transform we can compute $Q_n$, in $O(n \log n)$ time.

1. Append $n$ zeros to the end of $Q_1$ and $\pi_n$, forming $Q^{(n)}$ and $\pi^{(n)}$.
2. Compute the FFT $\mathcal{F}(Q^{(n)})$ and $\mathcal{F}(\pi^{(n)})$.
3. Apply the convolution theorem $Q^{(n)} = \mathcal{F}^{-1}(\mathcal{F}(Q^{(n)} \ast \pi^{(n)}))$, where * denotes cyclic convolution and is pointwise multiplication.
4. Compute the inverse Fourier transform of $Q^{(n)}$ to obtain the vector $Q_n$.

Repeating this procedure for all $n$ yields a total running time of $O(n \log n)$.

This is the fastest known algorithm for computing the two-sided crossing probability and the first to break the $O(n^3)$ barrier.

New one-sided $O(n^2)$ algorithm

In the one-sided case ($g < 0$ or $h > 1$) an even faster algorithm is possible. The joint density of the random vector of uniform order statistics is $\Pi_{i=1}^n (1-t_i)$.

Hence the one-sided variant of Eq. (2) is given by

$$P_r \{ g(U_i) < b_i, h(U_i) > d_i \}$$

Numerically evaluating this integral from right to left takes $O(n^2)$ time. A naive implementation fails at $n = 130$ due to numerical errors, but with some effort we have been able to get up to $n = 50$,000.

**Application:** p-value computation for goodness-of-fit statistics

The p-value of several sup-type continuous goodness-of-fit statistics directly translates to a probability of the form of Eq. (1). Hence we can compute such $p$-values in $O(n^2 \log n)$ time. The following table demonstrates that this improvement is not merely theoretical but yields a significant reduction in running times.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Two-sided</th>
<th>One-sided</th>
</tr>
</thead>
<tbody>
<tr>
<td>K&amp;S 2001</td>
<td>0.3 sec</td>
<td>12 sec</td>
</tr>
<tr>
<td>O(n^2 log n) algorithm</td>
<td>0.3 sec</td>
<td>12 sec</td>
</tr>
<tr>
<td>K&amp;S 2001</td>
<td>13 sec</td>
<td>18 minutes</td>
</tr>
<tr>
<td>O(n^2 log n) algorithm</td>
<td>13 sec</td>
<td>18 minutes</td>
</tr>
</tbody>
</table>

Table 1: Running times for computing $p$-values of the $M_i$ goodness-of-fit statistics of Berk & Jones.

Summary

- State-of-the-art $O(n^3 \log n)$ algorithm for computing the two-sided crossing probability of empirical CDFs and Poisson processes.
- Fast $O(n^2)$ algorithms for the one-sided case.
- Potential applications include: $p$-value and power calculations for goodness-of-fit statistics, construction of $\alpha$-level confidence bands for distribution functions, analysis of boundary crossing and first passage of a Brownian motion, queuing theory, sequential testing...
- Efficient C++ code at: [http://www.wisdom.weizmann.ac.il/~amitmo](http://www.wisdom.weizmann.ac.il/~amitmo)

References