מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

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 $\hat{F}_n$ 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

$$\Pr\left[\forall t : g(t) < \hat{F}_n(t) < h(t)\right] \tag{1}$$

Several algorithms have been proposed over the years, all are $O(n^3)$. [Epanechnikov 1968, Steck 1971, Noé 1972, Friedrich & Schellhaas 1998, Khmaladze & Shinjikashvili 2001].

Equivalent formulation

Let $U_{1:n} \leq U_{2:n} \leq \ldots \leq U_{n:n}$ be the order statistics of n draws from $U[0,1]$. Given arbitrary bounds $b_1, \ldots, b_n, B_1, \ldots, B_n \in \mathbb{R}$ compute the probability

$$\Pr\left[\forall i : b_i < U_{i:n} < B_i\right]. \tag{2}$$

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

Lemma. $\hat{F}_n$ satisfies $g(t) < \hat{F}_n(t) < h(t)$ for all t if and only if it satisfies these inequalities at all times when $n \cdot g(t)$ or $n \cdot h(t)$ cross an integer.

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

$$R(s,m) := \Pr\left[\forall t \in [0,s] : g(t) < \hat{F}_n(t) < h(t) \text{ and } \hat{F}_n(s) = \tfrac{m}{n}\right].$$

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

$$R(t_{i+1}, m) = \begin{cases} \sum_\ell R(t_i, \ell) \cdot \Pr\left[(t_i, \ell) \to (t_{i+1}, m)\right] & \text{if } g(t_{i+1}) < m/n < h(t_{i+1}) \\ 0 & \text{otherwise.} \end{cases}$$

where $\Pr\left[(t_i,\ell) \to (t_{i+1}, m)\right] = \Pr\left[\text{Binomial}(n-\ell, \frac{t_{i+1}-t_i}{1-t_i}) = m - \ell\right]$.
Solution. Eq. (1) is equal to $R(1,n)$, which can be computed in $O(n^3)$.

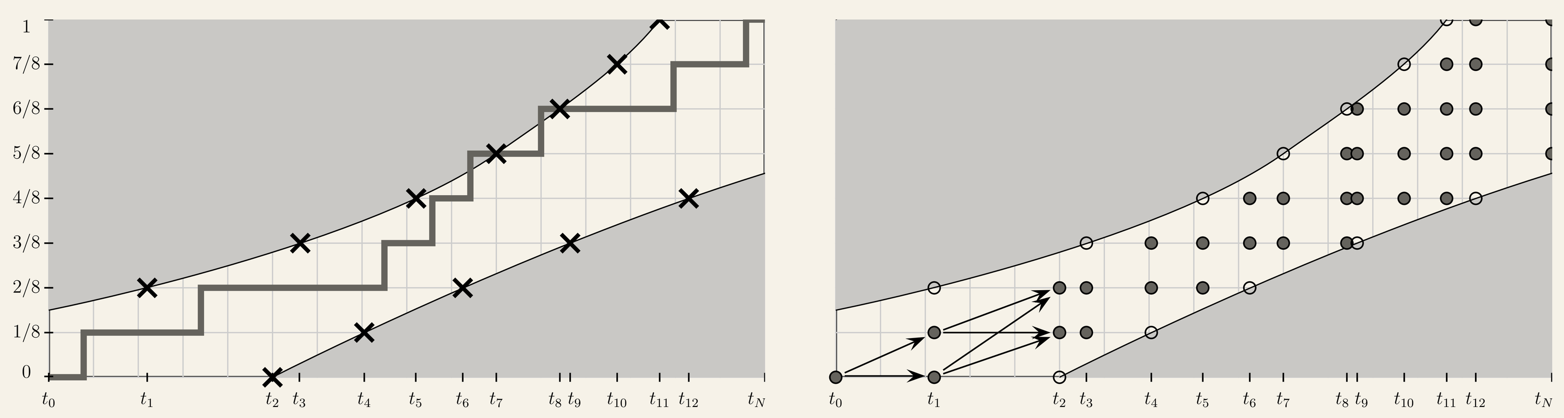

Figure 1: (left panel) **x** marks the i/n crossing points. $\hat{F}_n$ crosses one of the boundaries if and only if it crosses an **x** mark; (right panel) Layer graph representing the entries $R(t_i, m)$.

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

Lemma. The distribution of the stochastic process $n \cdot \hat{F}_n(t)$ is identical to that of a Poisson process $\xi_n(t)$ with intensity n conditioned on $\xi_n(1) = n$.

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

$$Q(s,m) := \Pr\left[\forall t \in [0,s] : g(t) < \tfrac{1}{n}\xi_n(t) < h(t) \text{ and } \xi_n(s) = m\right].$$

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

$$Q(t_{i+1}, m) = \begin{cases} \sum_\ell Q(t_i, \ell) \cdot \Pr\left[Z_i = m - \ell\right] & \text{if } g(t_{i+1}) < m/n < h(t_{i+1}) \\ 0 & \text{otherwise} \end{cases}$$

where Z_i is a Poisson random variable with intensity $n(t_{i+1} - t_i)$.
Solution. Apply the lemma to obtain

$$\Pr\left[\forall t : g(t) < \hat{F}_n(t) < h(t)\right] = Q(1,n) / \Pr\left[\text{Poisson}(n) = n\right].$$

Computing Q(1,n) still requires $O(n^3)$ operations.

New two-sided $O(n^2 \log n)$ algorithm

Denote $Q_{t_i} = (Q(t_i, 0), \ldots, Q(t_i, n))$ and $\pi_\lambda = (\Pr[Z_\lambda = 0], \ldots, \Pr[Z_\lambda = n])$ where Z_λ is a Poisson random variable with expected value λ.

Key idea: the vector $Q_{t_{i+1}}$ is nothing but a truncated linear convolution of Q_{t_i} and $\pi_{n(t_{i+1}-t_i)}$. Hence, using the circular convolution theorem and the Fast Fourier Transform we can compute $Q_{t_{i+1}}$ in $O(n \log n)$ time.

1. Append n zeros to the end of Q_{t_i} and $\pi_{n(t_{i+1}-t_i)}$, forming Q^{2n} and π^{2n}.
2. Compute the FFT $\mathcal{F}\{Q^{2n}\}$ and $\mathcal{F}\{\pi^{2n}\}$.
3. Apply the convolution theorem $C^{2n} = \mathcal{F}\{Q^{2n} \star \pi^{2n}\} = \mathcal{F}\{Q^{2n}\} \cdot \mathcal{F}\{\pi^{2n}\}$, where $\star$ denotes cyclic convolution and $\cdot$ is pointwise multiplication.
4. Compute the inverse Fourier transform of C^{2n} to obtain the vector $Q_{t_{i+1}}$

$$Q_{t_{i+1}}(m) = \begin{cases} \mathcal{F}^{-1}\{C^{2n}\}(m) & \text{if } g(t_{i+1}) < m/n < h(t_{i+1}) \\ 0 & \text{otherwise.} \end{cases}$$

Repeating this procedure for all i yields a total running time of $O(n^2 \log n)$. 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

$$f(U_{1:n}, \ldots, U_{n:n}) = \begin{cases} n! \text{ if } 0 \leq U_{1:n} \leq \ldots \leq U_{n:n} \leq 1, \\ 0 \text{ otherwise.} \end{cases}$$

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

$$\begin{aligned} \Pr\left[\forall i : b_i < U_{i:n}\right] &= n! Vol\{(U_{1:n}, \ldots, U_{n:n}) \mid \forall i : b_i < U_{i:n} \leq U_{i+1:n}\} \\ &= n! \int_{b_n}^{1} dU_{n:n} \int_{b_{n-1:n}}^{U_{n:n}} dU_{(n-1)} \ldots \int_{b_2}^{U_{3:n}} dU_{2:n} \int_{b_1}^{U_{2:n}} dU_{1:n}. \end{aligned}$$

Numerically evaluating this integral from right to left takes $O(n^2)$ time. A naïve implementation fails at $n \approx 150$ due to numerical errors, but with some effort we have been able to get up to $n \approx 50,000$. [3]

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.

Two-sided	$n = 4000$	$n = 16,000$	$n = 64,000$	$n = 256,000$
K&S 2001	0.5 sec	8 sec	94 sec	18 minutes
$O(n^2 \log n)$ algorithm	0.3 sec	2 sec	15 sec	117 sec
One-sided				
K&S 2001	45 sec	24 minutes	18 hours	weeks
$O(n^2 \log n)$ algorithm	2 sec	29 sec	9 minutes	3 hours
$O(n^2)$ algorithm	1.3 sec	19 sec	n/a	n/a

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

Summary

- State-of-the-art $O(n^2 \log n)$ algorithm for computing the two-sided crossing probability of empirical CDFs and Poisson processes.
- Fast $O(n^2)$ algorithm for the one-sided case.
- Potential applications include: p-value and power calculations for goodness-of-fit statistics, construction of α-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

References

[1] Thomas Friedrich and Helmut Schellhaas. Computation of the percentage points and the power for the two-sided Kolmogorov-Smirnov one sample test. *Statistical papers*, 39:361–375, 1998.

[2] Estate Khmaladze and Eka Shinjikashvili. Calculation of noncrossing probabilities for poisson processes and its corollaries. *Advances in applied probability*, 33:702–716, 2001.

[3] Amit Moscovich-Eiger, Boaz Nadler, and Clifford Spiegelman. On the exact Berk-Jones statistics and their p-value calculation. *ArXiv e-prints*, June 2015.

[4] Amit Moscovich-Eiger and Boaz Nadler. Fast calculation of boundary crossing probabilities for Poisson processes. *ArXiv e-prints*, March 2015.