How does null history length affect martingale detection of a Gaussian shift?

Oliver Hennhöfer · GPT-6 Astra

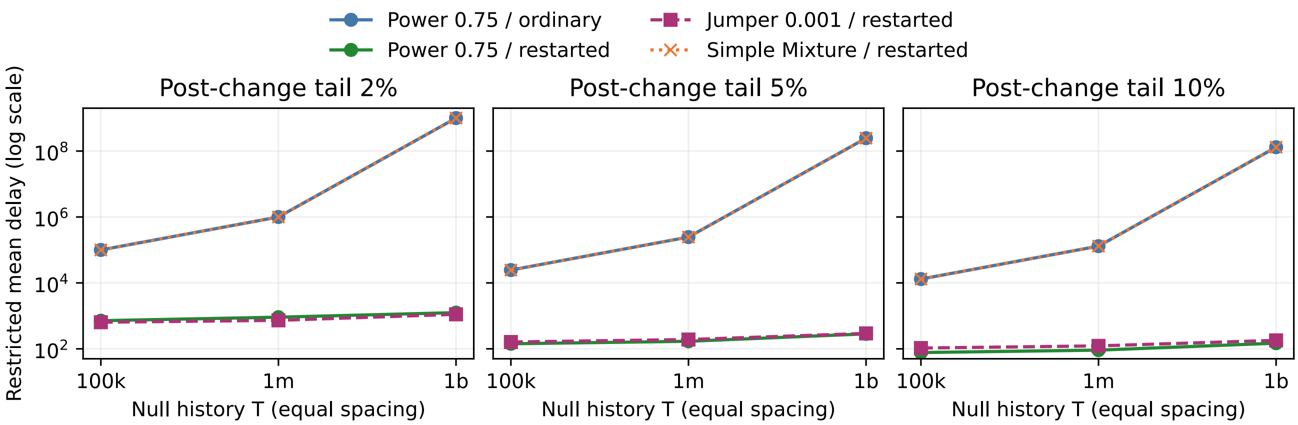
11 September 2026

Main finding. A long period without change can make a detector slow to react. **Restarted Power and Simple Jumper largely avoid this delay** in our simulation. Restarted Simple Mixture does not: it still carries the influence of the entire earlier history.

Experiment. Each sequence has T observations from `cal` $N(0, 1)$, followed by T from `cal` $N(\mu, 1)$, with $T =$ 100k, 1 million, or 1 billion. Mean shifts of about 0.273, 0.681, and 1.045 raise the chance of exceeding the original 99th percentile from 1% to 2%, 5%, and 10%. We use 32, 16, and 4 runs per condition at the respective lengths: 208 runs including no-change controls. Conditions share random draws for paired comparisons.

We calculate $p_t = \Phi(-X_t)$ from the known Gaussian baseline: the **oracle** setting. We test Power ($\varepsilon = 0.25, 0.5, 0.75$), Simple Jumper ($j = 0.001, 0.01, 0.1$), and an equal Power mixture ($\varepsilon = 0.1, 0.25, 0.5, 0.75, 1$). Each supplies ordinary, restarted-Ville, CUSUM, and Shiryaev–Roberts (SR) evidence at thresholds 20, 100, 1,000, and 1 million.

Reading the figure. Selected methods use threshold 20; **lower means faster detection**. We exclude runs that alarm before the change, then average the observations until detection, counting T for each undetected run. This is the **restricted mean delay**: missed changes count as a full window, not zero. The vertical scale is logarithmic.



Results. For the 5% tail condition, ordinary Power 0.75 needs about 24,600, 246,000, and 246 million observations as history grows; **restarted Power needs only 142, 170, and 285**. Every run avoids an early alarm and detects. Restarted Simple Mixture remains almost as slow as ordinary Power. For the weaker 2% tail condition, neither detects within the window at any history length. Restarted Power detects in every run, taking about 715, 914, and 1,260 observations. Restarted Jumper 0.001 detects all shifts in every run without an early alarm: 29/32, 14/16, and 4/4 runs, respectively. Its averages exclude those early false alarms.

Why this happens. Ordinary Power usually loses betting capital during a stable period and must recover it before triggering an alarm. A longer history creates a larger deficit. Restarted-Ville evidence gives recent observations a fresh start, with less weight on later starts. Simple Mixture's neutral component gains most of the betting weight during the stable period. **Restarting its alarm calculation does not reset those weights.**

Why CUSUM/SR look weak. At threshold 20 and $T = 1$ billion, every SR run and every CUSUM run using an individual Power/Jumper rule alarms before the change. **They look powerless because they alarm too early.** Zero first alarms afterward does not establish poor sensitivity. The same threshold does not give equal protection: valid Ville statistics have a 5% lifetime false-alarm bound at 20; CUSUM/SR do not.

Limits. Four runs per condition at $T = 1$ billion give only rough evidence about detection probability. **These are known-Gaussian results, not long-history distribution-free conformal results.** The conformal implementation has only two short test runs and fails the projected billion-run memory budget. Oracle storage is bounded by chunk size; peak memory was not measured.