Repeated Significance Tests under Budget Constraints

Vladimir Pozdnyakov

University of Connecticut

2015

Repeated Significance Test (RST)

Let $\{T_n\}_{1\leq n\leq N}$ be a sequence of test statistics (based on a data set of size n) that we want to employ to test some H_0 versus a certain alternative H_a.

- **Traditional approach**

  We reject H_0 if $|T_N|/A_N$ is large. Here A_N is an appropriate (perhaps random) normalizing sequence.

- **Sequential approach**

  Let n_0 be an initial sample size, and N is a target sample size. Consider the following test statistic

  $$h_N = \max\left\{\frac{|T_n|}{A_n} : n_0 \leq n \leq N\right\}.$$

  We reject H_0 if the value of the test statistic h_N is too large. Note that if $|T_n|/A_n$ is already large for $n < N$ we can stop and make a decision.

Invariance Principle

Mathematically, the sequential testing is often a boundary crossing problem, where $\{T_n\}_{1\leq n\leq N}$ is a trajectory, and $\{A_n\}_{1\leq n\leq N}$ is a boundary.

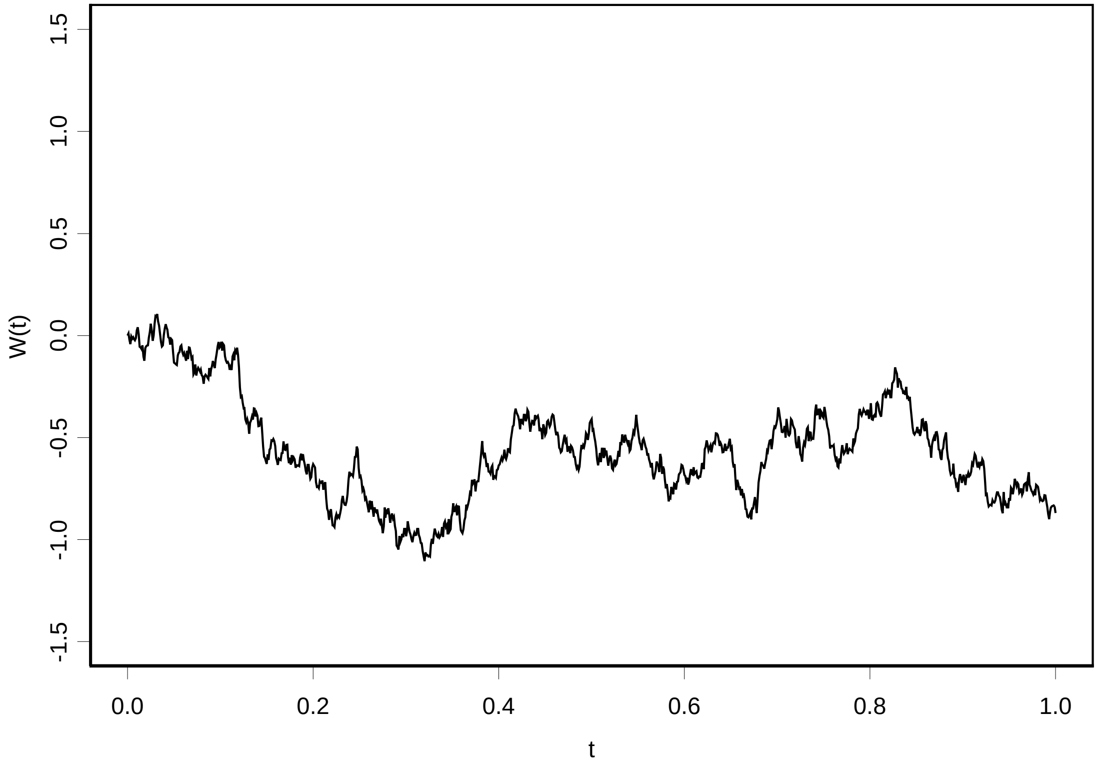
One can introduce a stoping time

$$\tau = \mathrm{inf}\{n : n_0 \leq k \leq N, |T_n| > |A_n|\},$$

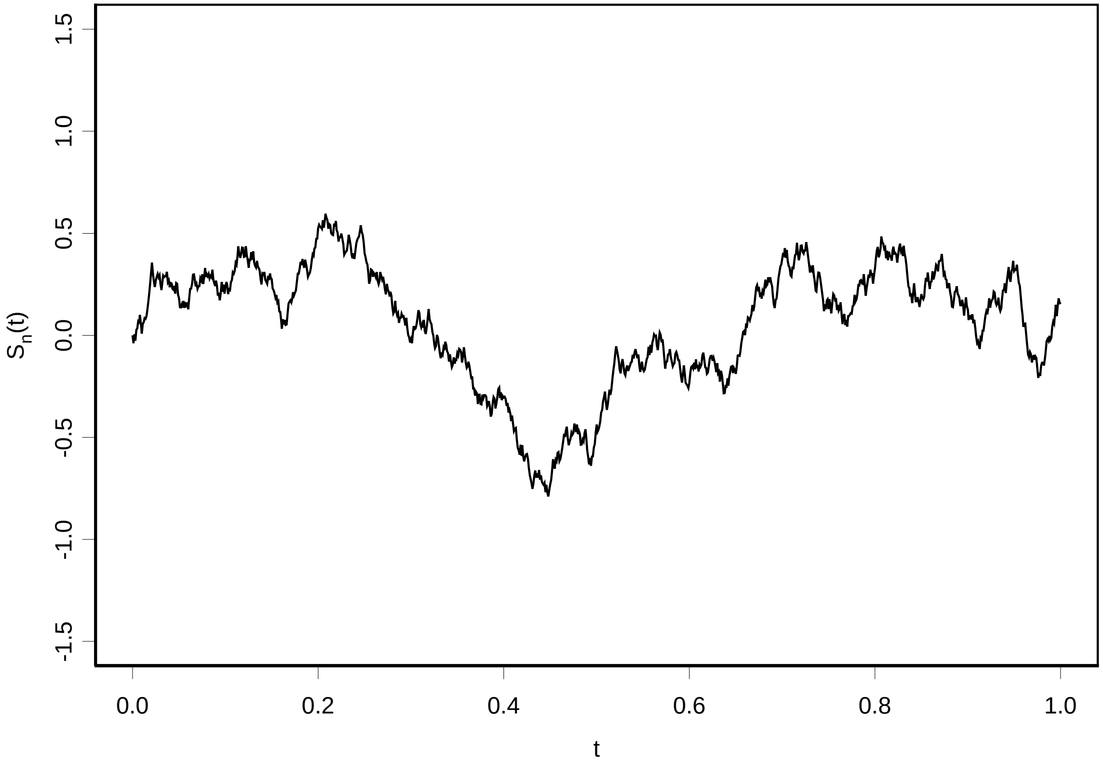
and then reject H_0 whenever $\tau < N$.

Therefore, if $\{T_n\}_{1\leq n\leq N}$ can be approximated by a well-known process (for instance, Brownian Motion), most likely the solution of the boundary crossing problem is available.

Brownian Motion

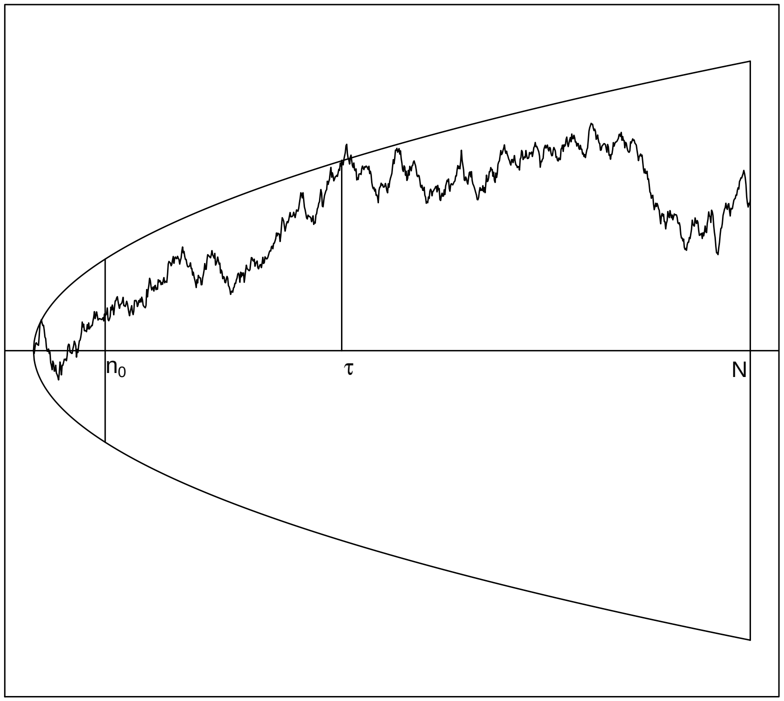


Classical Invariance Principle



Uniform $U[-10, 10]$ random walk normalized by its sample standard deviation, $n = 1000$

Repeated Significance Test

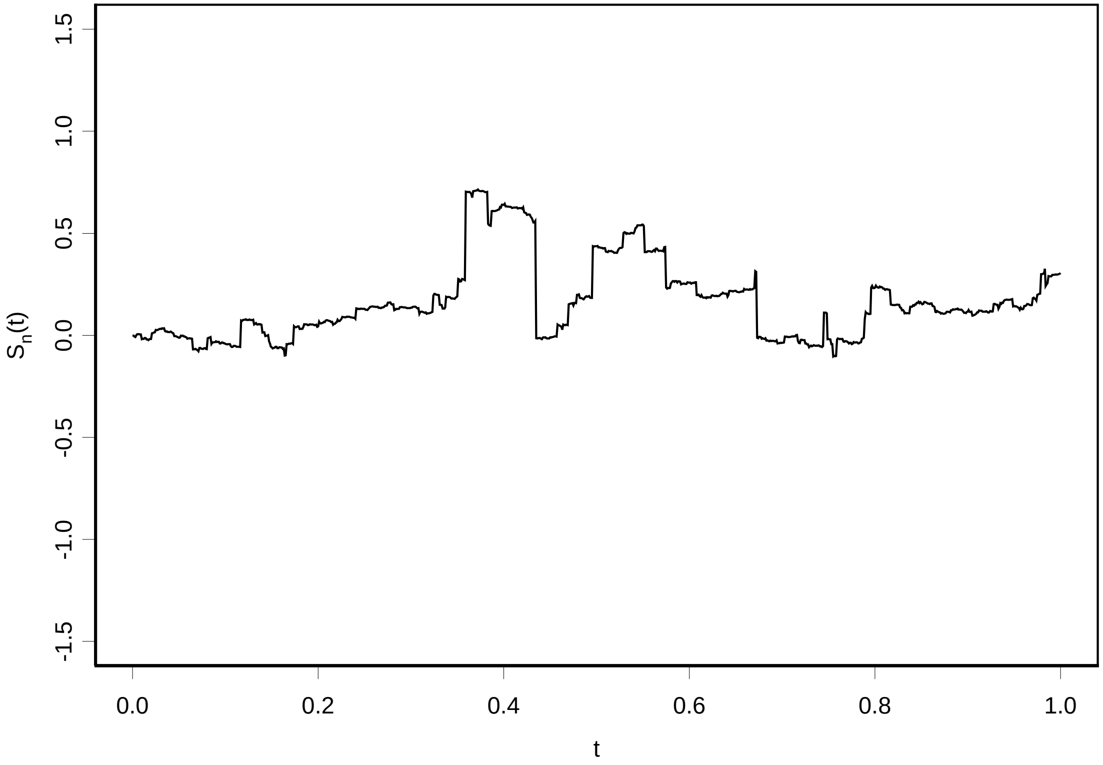


Random Stopping $\mathcal{N}$

Two examples when it is beneficial to introduce random target sample size $\mathcal{N}$ instead of deterministic N:

- Heavy-tailed Distributions: Adaptive Target Sample Size.
- Budget Constraints on Sampling Procedure

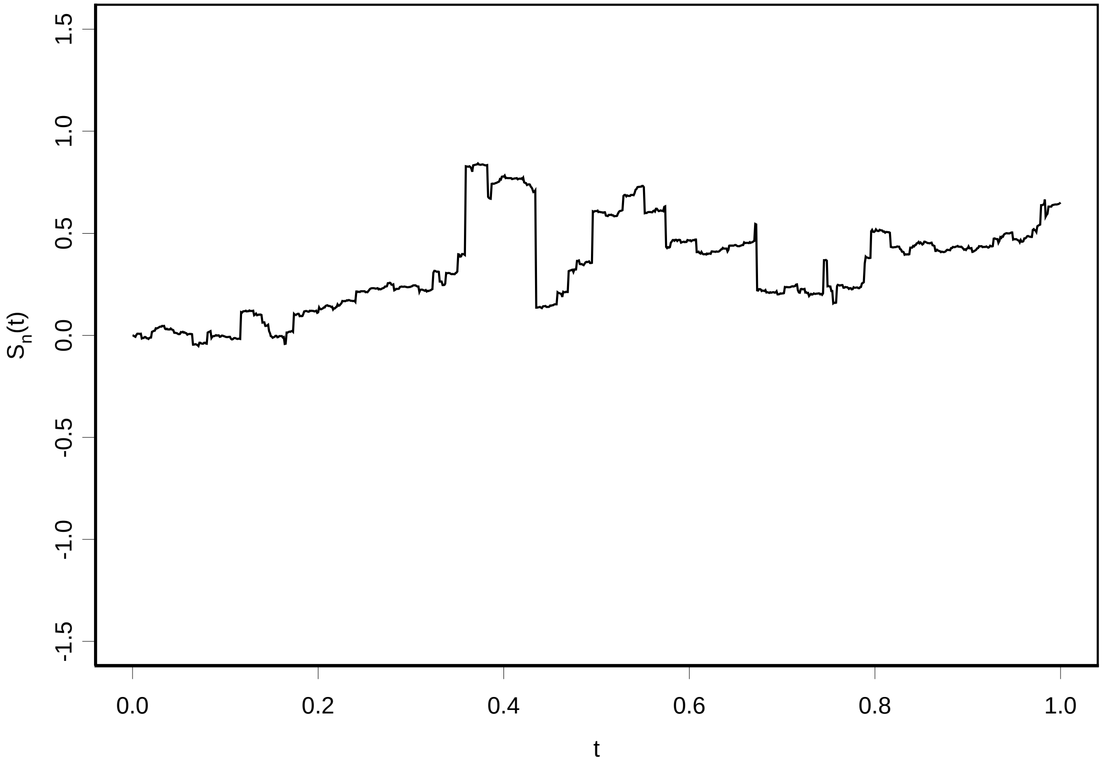
Heavy-tailed Distributions



Cauchy random walk normalized

by its sample standard deviation, $n = 1000$

Heavy-tailed Distributions



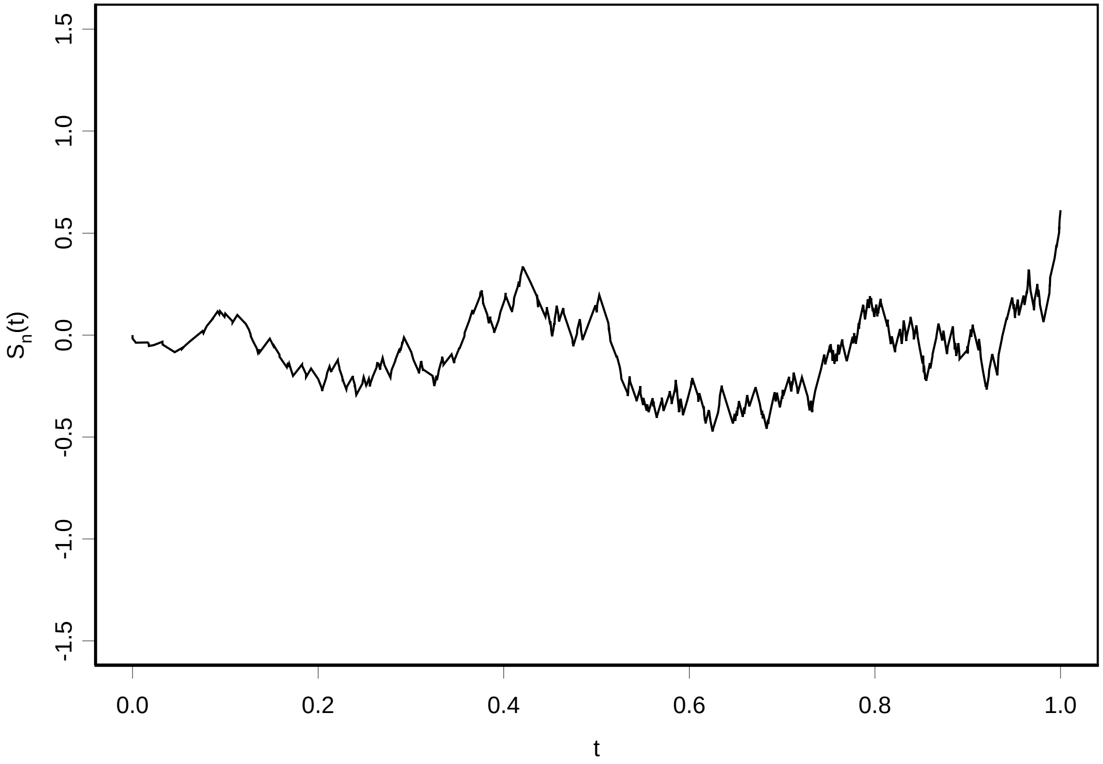
Non-centered ($\mu = .2$) Cauchy random walk normalized by its sample standard deviation, $n = 1000$

Appropriate Invariance Principle

Cauchy Random Walk does not converge to anything.

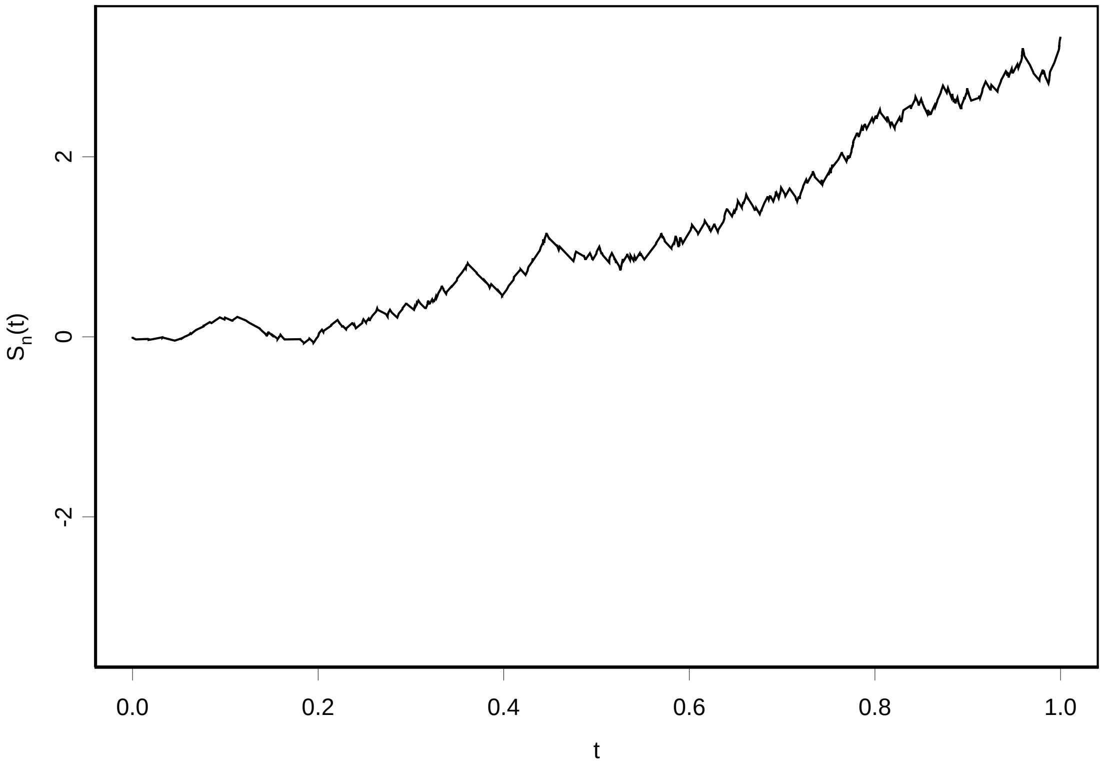
However, if it is truncated in a certain way it can be approximated by the Brownian motion again.

Truncated Random Walk



Truncated Cauchy random walk normalized by its sample standard deviation, $n = 1000$

Truncated Random Walk



Truncated non-centered ($\mu = .2$) Cauchy random walk normalized by its sample standard deviation, $n = 1000$

Difficulty

The critical value for test statistic

$$h_N = \max\left\{\frac{|T_n|}{A_n} : n_0 \le n \le N\right\}.$$

for the classical RST depends on significance level α and ratio n_0/N.

In the heavy tail case (Glaz and Pozdnyakov (2005)), we also need to know the tail index, because the shape of boundary A_n depends on it (in classical case it's usually constant$\times\sqrt{n}$).

But what if we do not know it?

A Solution: Adaptive Target Sample Size

Define a stopping time $\mathcal{N}$ by

$$\mathcal{N} = \text{inf}\left\{k \geq n_0 : \frac{A_k}{A_{n_0}} \geq \frac{1}{t_0}\right\},$$

and use $h_{\mathcal{N}}$ instead of h_N (see Glaz and Pozdnyakov (2007)).

The test will stop itself when it's time to stop.

The budget constraint connection:

$$\mathcal{N} = \text{inf}\left\{k \geq n_0 : A_k \geq \frac{1}{t_0} A_{n_0}\right\},$$

Sensor Networks

For many tasks of reconnaissance and surveillance, networks of spatially distributed sensors provide the low-cost, low-risk solution, and the rapidly growing field of distributed sensing now provides many interesting challenges for probabilistic modeling (see, e.g. Chong and Kumar (2003)). Here we examine a simple model that joins the rudiments of sequential decision making with the engineering constraint of a fixed budget for the cost of the transmissions sent to and from the distributed sensors. The costs associated with transmissions from the sensors are intended to either real physical expenditures, such as battery life, or to capture more subtle costs, such as the cumulative risk of a remote sensor (or bug) being found by an adversary.

Decentralized Architecture

Our focus is specifically on distributed sensor networks with a *decentralized architecture*; that is, each sensor is capable of certain preliminary decision processing before sending summarized information to an automated *fusion agent*. The fusion agent is thought of as being remote from the distributed sensors, and it has two responsibilities. First, it interrogates the distributed sensors according to some protocol, and, second, it makes an "over all" network decision at such as time as sufficient evidence has been collected to support a decision. In the protocols considered here, the fusion agent remotely interrogates the distributed sensors sequentially one-by-one.

Cost and Budget

To fix notation, we first describe a version of our problem that is a bit more general than the one we analyze in detail. By $\{X_i\}_{i\geq 1}$ we denote a sequence of independent identically distributed random variables that we view as observations coming from a sequence of queries that are put to the distributed sensors. Next, we consider a nonnegative function $c(\cdot)$ and we view $c(X_i)$ as the cost of the fusion agent collecting the value X_i. We then let B denote our budget for payment for the costs of this collected information.

Decision Rule

We assume that we can continue to collect distributed sensor information so long as our cumulative cost C_n satisfies the budget constraint

$$C_n \overset{\text{def}}{=} c(X_1) + c(X_2) + \cdots + c(X_n) < B,$$

and we assume that decision rule for the fusion agent is based on the sum of the observations

$$S_n = X_1 + X_2 + \cdots + X_n.$$

A decision is made by the fusion agent at the time when the process $\{S_n\}$ hits either an upper boundary $\mathcal{U} \equiv \{U(n) : n = 1, 2, ...\}$, say for a positive decision, or a lower boundary $\mathcal{L} \equiv \{L(n) : n = 1, 2, ...\}$ for a negative decision. Naturally, the upper and lower boundaries depend on the hypothesis that the fusion must test to make the network decision. In general, these boundaries are curved, and they would be determined by the usual tools of sequential decisions theory such as the sequential probability ratio test (or its extensions and approximations).

Boundary Crossing Problem

Thus, our distributed sensing problem leads to a special kind of boundary crossing problem for the two-dimensional process $Z_n = (C_n,\, S_n)$, $n = 1, 2, \ldots$. From the design prospective, the random variables of most interest are the *decision time*,

$$\tau_D = \text{inf}\{n : S_n \geq U(n) \ \text{ or } \ S_n \leq L(n)\,\}, \tag{1}$$

and the *budget exhaustion time*,

$$\tau_E = \text{inf}\{n : C_n \geq B\,\}. \tag{2}$$

Large Sample Size Case

This frames the general problem, but without further specialization, the tools for analysis are limited. Guerriero at el. (2010) considered the case when B is large and used the renewal theorem approximation $\tau_E \sim B/\mathbf{E}(c(X_1))$ to make some progress.

More specifically, the LLN for τ_E, the classical invariance principle and the mapping theorem can be employed to construct a sequential procedures.

However, a boundary crossing problem with small B looks interesting and challenging, because in this case we need exact formulas.

Trinomial Response

In Steele and Pozdnyakov (2010) we are mostly concerned with a “binary-plus-passive” design where at the time a sensor is interrogated it checks its locally stored observation data and acts according to a three part rule: (a) if the observation data has a “large weight” (in some appropriate sense) it sends $+1$ to the fusion center, (b) if it has a correspondingly “small weight” then the sensor sends -1, and, finally, (c) if the observation is within a certain intermediate range, the sensor *does not reply at all*. Here the fusion agent *does* know that the sensor was interrogated, so the non-reply conveys useful information. The benefit of this “passive reply” alternative is that the sensor does not expend energy or expose itself to incremental risk of adversary detection. Naturally, we view an active response from a sensor as expensive while we view a non-response as relatively cheap — but not entirely costless.

More Specifically

To have any hope for exact formulas, one needs more detailed information on distribution of the X_i, the cost function c, and the decision boundaries. The simplest non-trivial case begins with a trinomial model for the X_i that we parameterize as $\mathbf{P}(X_i = 1) = p$, $\mathbf{P}(X_i = 0) = r$, and $\mathbf{P}(X_i = -1) = q$ with non-negative p, r, q such that $p + r + q = 1$. We then take the simplest possibilities for the decision boundaries; for the upper boundary $\mathcal{U}$ we take a constant U and for the lower boundary $\mathcal{L}$ we take the constant $-L$ where $U > 0$ and $L > 0$ are integers.

To capture the benefit of the "non-response" possibility in the binary-plus-passive protocols we introduce an integer K and to consider the cost function defined by

$$c(x) = |x| + \delta(x)/K \tag{3}$$

where and $\delta(0) = 1$ and $\delta(x) = 0$ if $x \neq 0$. In other words, our cost for the binary transmission of 1 or -1 from a distributed sensor is a "unit", and the cost of the passive transmission from a distributed sensor (i.e. a non-response to a query) is just a $1/K$ fraction of a "unit".

Decision and Exhaustion Times

The decision time (1) is now more explicitly given by

$$\tau_D = \text{inf}\{n : S_n \geq U \text{ or } S_n \leq -L\}, \tag{4}$$

and the budget exhaustion time is given by (3) and

$$\tau_E = \text{inf}\{n : c(X_1) + c(X_2) + \cdots + c(X_n) \geq B\}. \tag{5}$$

Here one should note that the total cost C_{τ_E} at the time the budget is exceeded can take on the any of the values $B, B + 1/K, ..., B + (K-1)/K$, but an "overshoot" of the budget is only possible if the sensor response at time τ_E was an active response.

Probabilities of Interest

There are several probabilities that can inform us about the design of a binary-plus-passive distributed sensor network. We are particularly interest in

$$\mathbf{P}(\tau_D \leq \tau_E), \tag{6}$$

the probability that we make a decision before we over run our budget. We are also interested in this event together with the kind of decision that we make:

$$\mathbf{P}(\tau_D \leq \tau_E \text{ and } S_{\tau_D} = U) \quad \text{and} \quad \mathbf{P}(\tau_D \leq \tau_E \text{ and } S_{\tau_D} = -L) \tag{7}$$

Moreover, it may be useful sometimes to know just where the evidence stands when no decision has been made and yet the budget is exhausted; this is given by

$$\mathbf{P}(\tau_E < \tau_D, \text{ and } S_{\tau_E} = x). \tag{8}$$

Associated Binomial Random Walk

A $\{-1, 0, 1\}$ trinomial process can be associated with binomial process simply by deleting zeros and the times at which they occur. This process of "casting out zeros" distorts the values (and distribution) of hitting times, but it does so in way that still permits useful calculations. To make this explicit, we first introduce a new i.i.d. sequence $\{X_i^* : i = 1, 2, \ldots\}$ with

$$\mathbf{P}(X_i^* = 1) = p_* \text{ and } \mathbf{P}(X_i^* = -1) = q_*, \text{ where}$$

$$p_* = p/(p+q) \text{ and } q_* = q/(p+q),$$

and we consider the new binomial random walk $S_n^* = X_1^* + X_2^* + \cdots + X_n^*$ together with a corresponding "decision time",

$$\tau_D^* = \inf\{n : S_n^* \geq U \text{ or } S_n^* \leq -L\}. \tag{9}$$

Finally, it will be useful in our analysis to consider the number ν of active responses observed up to and including time τ_E. Under our binary-passive protocol this is simply

$$\nu = \sum_{i=1}^{\tau_E} |X_i|. \tag{10}$$

Main Result

Theorem 1 *For each* $0 \le n \le B$*, one has*

$$\mathbf{P}(\nu = n) = \binom{(B-n)K+n}{n}(1-r)^n r^{(B-n)K} + \sum_{i=1}^{K-1} \binom{n-1+(B-n)K+i}{n-1}(1-r)^n r^{(B-n)K+i}.$$

Moreover,

$$\mathbf{P}(\tau_D > \tau_E) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* > n)\,\mathbf{P}(\nu = n),$$

$$\mathbf{P}(\tau_D \le \tau_E, S_{\tau_D} = U) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* \le n, S^*_{\tau_D^*} = U)\,\mathbf{P}(\nu = n),$$

$$\mathbf{P}(\tau_D \le \tau_E, S_{\tau_D} = -L) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* \le n, S^*_{\tau_D^*} = -L)\,\mathbf{P}(\nu = n), \text{ and}$$

$$\mathbf{P}(\tau_D > \tau_E, S_{\tau_E} = x) = \sum_{n=0}^{B} \mathbf{P}(\tau_D^* > n, S_n^* = x)\,\mathbf{P}(\nu = n), \text{ where } -L < x < U.$$

Main Result

Theorem 1 tells us how the basic probability results for τ_D, τ_E, S_{τ_D}, and S_{τ_E} can be expressed in terms of the more easily analyzed (or well known) quantities ν, τ_D^*, τ_E^*, $S_{\tau_D}^*$, and $S_{\tau_E}^*$. One should note that given the first formula, all of the terms on the right side of the subsequent formulas can be readily computed since the all "starred" variables refer to the standard biased random walk $\{S_k^* : k = 0, 1, \ldots\}$ for which formulas (or methods) for all of the required probabilities are well-known.

Numerical Comparisons

For the renewal theory approximations to have a fighting chance, the budget B must be of at least moderate size, so we first consider cases $B = 20$. To specify the rest of the model, we take the signal probabilities to be $p = 1/10$ and $q = 1/10$ (so $r = 8/10$), take the decision limits to be $U = 10$ and $L = 10$, and, finally, take the cost to send a passive response to be $1/K = 1/8$

	$\mathbf{P}(\tau_D \leq \tau_E)$		
Exact	.0113587		
Approximation	Round	Floor	Ceiling
Trinomial	.0138173	.0138173	.0130238
Binomial	.00683594	.00683594	.0147705

For the second example we decrease B just a little to 17. We then take $p = 1/5$, $q = 1/5$, $r = 3/5$, $U = 8$, $L = 8$, and $K = 20$ to complete the model.

	$\mathbf{P}(\tau_D \leq \tau_E)$		
Exact	.0971227		
Approximation	Round	Floor	Ceiling
Trinomial	.0930403	.0876269	.0930403
Binomial	.0980835	.0703125	.0980835

References

- C.-X. Chong and S. P. Kumar, Sensor Networks: Evolution, Opportunities, and Challenges, *Proc. IEEE*, 91 (2003), 1247-1256.

- J. Glaz and V.I. Pozdnyakov, A Repeated Significance Test for Distributions with Heavy Tails, *Sequential Analysis*, 24 (2005), 77-98.

- V. Pozdnyakov and J. Glaz , A Nonparametric Repeated Significance Test with Adaptive Target Sample Size, *Journal of Statistical Planning and Inference*, 137 (2007), 869-87

- M. Guerriero, V. Pozdnyakov, J. Glaz, and P. Willett, A Repeated Significance Test with Applications to Sequential Detection in Sensor Networks, *IEEE Transactions on Signal Processing*, 58 (2010), 3426-3435.

- V. Pozdnyakov and J.M. Steele, Bugs on a Budget: Distributed Sensing with Cost for Reporting and Non-Reporting, *Probability in the Engineering and Informational Sciences*, 24 (2010), 525-534.

- A. Gut, The gambler's ruin with delays, *Statistics & Probability Letters*, 83 (2013), 2549-2552.

- A. Gut, Some remarks on distributed sensing with costs. *Probability in the Engineering and Informational Sciences*, 28 (2014), 271-277.