Sequential Hypothesis Testing and Changepoint Detection: Past and Future

Alexander G. Tartakovsky

University of Southern California, Los Angeles, CA USA

1. Nearly Optimal Sequential Tests of Composite Hypotheses. Let X_1, $X_2, \dots$ be a sequence of independent and identically distributed (iid) observations, and let $p_\theta(x)$ be a density (parametrized by a parameter θ) with respect to some non-degenerate sigma-finite measure μ. In his 1947 book, Wald [9, Section 6] suggested two approaches for modifying the *Sequential Probability Ratio Test* (SPRT) to test a simple null hypothesis $H_0 : \theta = \theta_0$ against a composite alternative $H_1 : \theta \in \Theta_1$. One method is to replace the likelihood ratio (LR) $\Lambda_n^\theta = \prod_{k=1}^n [p_\theta(X_k)/p_{\theta_0}(X_k)]$ by a weighted LR $\bar{\Lambda}_n = \int_{\Theta_1} w(\theta)\Lambda_n^\theta \, \mathrm{d}\theta$, using a suitably selected weight function $w(\theta)$ on the hypothesis H_1. This leads to the *Weighted SPRT* (WSPRT) $\bar{\delta} = (\bar{T}, \bar{d})$ with the stopping time $\bar{T}(A_0, A_1) = \inf\left\{n \geq 1 : \bar{\Lambda}_n \notin (A_0, A_1)\right\}$, $0 < A_0 < 1$, $A_1 > 1$. The weighted-based tests are also often called *mixture-based* tests of simply *mixtures*. The other way is to apply the generalized likelihood ratio (GLR) approach of classical fixed-sample size theory, employing the GLR statistic $\hat{\Lambda}_n = \sup_{\theta\in\Theta_1} \Lambda_n^\theta$ in place of the LR Λ_n^θ with *a priori* fixed parameters, which leads to the *Generalized Sequential Likelihood Ratio Test* (GSLRT) $\hat{\delta} = (\widehat{T}, \hat{d})$ with the stopping time $\widehat{T}(A_0, A_1) = \inf\left\{n \geq 1 : \hat{\Lambda}_n \notin (A_0, A_1)\right\}$.

In a more general case where the null hypothesis is also composite, $H_0 : \theta \in \Theta_0$, Wald [9] proposed to exploit the WSPRT with the weighted LR

$$\bar{\Lambda}_n = \frac{\int_{\Theta_1} w_1(\theta) \prod_{k=1}^n p_\theta(X_k) \, \mathrm{d}\theta}{\int_{\Theta_0} w_0(\theta) \prod_{k=1}^n p_\theta(X_k) \, \mathrm{d}\theta}.$$

Changing the measures and applying the Wald likelihood ratio identity, we obtain the upper bounds on the average error probabilities: $\bar{\alpha}_0(\bar{\delta}) = \int_{\Theta_0} \mathsf{P}_\theta(\bar{d} = 1) w_0(\theta) \, \mathrm{d}\theta \leq 1/A_1$, $\bar{\alpha}_1(\bar{\delta}) = \int_{\Theta_1} \mathsf{P}_\theta(\bar{d} = 0) w_1(\theta) \, \mathrm{d}\theta \leq A_0$. Clearly, for practical purposes, one would strongly prefer to upper-bound not the average error probabilities, which depend on a particular choice of weights, but rather the maximal error probabilities of Type I and Type II, i.e., to consider the class of tests $\mathbf{C}(\alpha_0, \alpha_1) = \{\delta : \sup_{\theta\in\Theta_0} \mathsf{P}_\theta(d = 1) \leq \alpha_0, \sup_{\theta\in\Theta_1} \mathsf{P}_\theta(d = 0) \leq \alpha_1\}$, $\alpha_0 + \alpha_1 < 1$. However, in general it is not clear how to obtain the upper bounds on maximal error probabilities of the WSPRT and the GSLRT. In this respect, the tests that are based on one-stage delayed estimators, for the first time suggested by Robbins and Siegmund [6, 7] in the context of power one tests in the beginning of seventies, represent a useful alternative considered below.

More generally, consider the following continuous- or discrete-time scenario with multiple composite hypotheses. Let $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathsf{P}_\theta)$, $t \in \mathbb{Z}_+ = \{0, 1, \dots\}$ or $t \in$

Author's email: tartakov@usc.edu

$\mathbb{R}_+ = [0, \infty)$, be a filtered probability space with standard assumptions about monotonicity and, in the continuous time case $t \in \mathbb{R}_+$, also right-continuity of the σ-algebras $\mathscr{F}_t$. The parameter $\theta = (\theta_1, \ldots, \theta_\ell)$ belongs to a subset $\tilde{\Theta}$ of ℓ-dimensional Euclidean space $\mathbb{R}_\ell$. The sub-σ-algebra $\mathscr{F}_t = \mathscr{F}_t^X = \sigma(\mathbf{X}_0^t)$ of $\mathscr{F}$ is generated by the stochastic process $\mathbf{X}_0^t = \{X(u),\ 0 \le u \le t\}$ observed up to time t . The hypotheses to be tested are "$H_i :\ \theta \in \Theta_i$", $i = 0, 1, \ldots, N$ $(N \ge 1)$, where Θ_i are disjoint subsets of $\tilde{\Theta}$. We will also suppose that there is an *indifference* zone $\mathrm{I}_{\mathrm{in}} \in \tilde{\Theta}$ in which there are no constraints on the probabilities of errors imposed. The indifference zone, where any decision is acceptable, is usually introduced keeping in mind that the correct action is not critical and often not even possible when the hypotheses are too close, which is perhaps the case in most, if not all, practical applications. However, in principle I_{in} may be an empty set. The probability measures P_θ and $\mathsf{P}_{\tilde{\theta}}$ are assumed to be locally mutually absolutely continuous, i.e., the restrictions P_θ^t and $\mathsf{P}_{\tilde{\theta}}^t$ of these measures to the sub-σ-algebras $\mathscr{F}_t$ are equivalent for all $0 \le t < \infty$ and all $\theta, \tilde{\theta} \in \tilde{\Theta}$.

A multihypothesis sequential test δ consists of the pair (T, d), where T is a stopping time with respect to the filtration $\{\mathscr{F}_t\}_{t \ge 0}$, and $d = d_T(\mathbf{X}_0^T) \in \{0, 1, \ldots, N\}$ is an $\mathscr{F}_T$-measurable (terminal) decision rule specifying which hypothesis is to be accepted once observations have stopped (the hypothesis H_i is accepted if $d = i$ and rejected if $d \ne i$, i.e., $\{d = i\} = \{T < \infty,\ \delta \text{ accepts } H_i\}$). The quality of a sequential test is judged on the basis of its error probabilities and expected sample sizes (or more generally on the moments of the sample size). Let $\alpha_{ij}(\delta, \theta) = \mathsf{P}_\theta(d = j) 1_{\{\theta \in \Theta_i\}}$ $(i \ne j,\ i, j = 0, 1, \ldots, N)$ be the probability of accepting the hypothesis H_j by the test δ when the true value of the parameter θ is fixed and belongs to the subset Θ_i. Introduce the class of tests $\mathbf{C}(||\alpha_{ij}||) = \left\{\delta : \sup_{\theta \in \Theta_i} \alpha_{ij}(\delta, \theta) \le \alpha_{ij},\ i, j = 0, 1, \ldots, N,\ i \ne j\right\}$ for which maximal error probabilities do not exceed the given numbers α_{ij}.

While almost all results hold for continuous time too, we will focus only on the discrete time scenario. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of θ. If in density $p_\theta(X_k)$ for the k^{th} observation we replace the parameter by the estimate $\hat{\theta}_{k-1}$ built upon the sample $(X_1, \ldots, X_{k-1})$ that includes $k-1$ observations, then $p_{\hat{\theta}_{k-1}}(X_k)$ is still a viable probability density, in contrast to the case of the GLR approach where $p_{\hat{\theta}_n}(X_k)$ is not a probability density anymore for $k \le n$. Therefore, the statistic

$$\Lambda_n^*(\theta_i) = \prod_{k=1}^{n} \frac{p_{\hat{\theta}_{k-1}}(X_k)}{p_{\theta_i}(X_k)} = \Lambda_{n-1}^*(\theta_i) \times \frac{p_{\hat{\theta}_{n-1}}(X_n)}{p_{\theta_i}(X_n)} \tag{1}$$

is a viable likelihood ratio, and it is the nonnegative P_{θ_i}-martingale with unit expectation, since $\mathsf{E}_{\theta_i}[\Lambda_n^*(\theta_i) | \mathbf{X}_1^{n-1}] = \Lambda_{n-1}^*(\theta_i)$. Therefore, one can use Wald's likelihood ratio identity for finding bounds on error probabilities if $\Lambda_n^*(\theta_i)$ is used instead of the LR with the true parameter value θ. Because of exactly this very convenient property as well as of the simple recursive structure (1) the hypothesis tests based on the adaptive LR's with one-stage delayed estimators represent a very attractive alternative to the GLR tests as well to the mixture-based tests. De-

fine the statistics $\Lambda_n^*(\Theta_i) = \prod_{k=1}^n p_{\hat{\theta}_{k-1}}(X_k)/\sup_{\theta\in\Theta_i}\prod_{k=1}^n p_\theta(X_k)$, $i = 0, 1, \ldots, N$. The multihypothesis test, which we will refer to as the *Multihypothesis Adaptive Sequential Likelihood Ratio Test* (MASLRT), has the form

$$\text{stop at the first } n \geq 1 \text{ such that for some } i \quad \Lambda_n^*(\Theta_j) \geq A_{ji} \quad \text{for all } j \neq i$$

and accept the (unique) H_i that satisfies these inequalities.

Write $\alpha_{ij}^*(\theta) = \mathsf{P}_\theta(d^* = j)1_{\{\theta\in\Theta_i\}}$ for the error probabilities of the MASLRT. It can be shown that $\sup_{\theta\in\Theta_i}\alpha_{ij}^*(\theta) \leq 1/A_{ij}$, $i \neq j$, so that $A_{ij} = 1/\alpha_{ij}$ implies $\delta^* \in \mathbf{C}(||\alpha_{ij}||)$.

For $r > 0$, the random variable ξ_n is said to converge P-r-*quickly* to a constant C if $\mathsf{E}L_\varepsilon^r < \infty$ for all $\varepsilon > 0$, where $L_\varepsilon = \sup\{n : |\xi_n - C| > \varepsilon\}$ ($\sup\varnothing = 0$).

Write $\lambda_n(\theta,\tilde{\theta}) = \log\frac{\mathrm{d}\mathsf{P}_\theta^n}{\mathrm{d}\mathsf{P}_{\tilde{\theta}}^n} = \sum_{k=1}^n \log\frac{p_\theta(X_k|\mathbf{X}_1^{k-1})}{p_{\tilde{\theta}}(X_k|\mathbf{X}_1^{k-1})}$ for the log-likelihood ratio (LLR) process. Assume that there exist positive and finite numbers $I(\theta,\tilde{\theta})$ such that

$$\frac{1}{n}\lambda_n(\theta,\tilde{\theta}) \xrightarrow[n\to\infty]{\mathsf{P}_\theta-r-\text{quickly}} I(\theta,\tilde{\theta}) \quad \text{for all} \quad \theta,\tilde{\theta}\in\Theta, \theta\neq\tilde{\theta}. \tag{2}$$

In addition, we certainly need some conditions on the behavior of the estimate $\hat{\theta}_n$ for large n, which should converge to the true value θ in a proper way. To this end, we require the following condition on the adaptive LLR process:

$$\frac{1}{n}\log\Lambda_n^*(\tilde{\theta}) \xrightarrow[n\to\infty]{\mathsf{P}_\theta-r-\text{quickly}} I(\theta,\tilde{\theta}) \quad \text{for all} \quad \theta,\tilde{\theta}\in\Theta, \theta\neq\tilde{\theta}, \tag{3}$$

so that the normalized by n LLR tuned to the true parameter value and its adaptive version converge to the same constants. In certain cases, but not always, conditions (2) and (3) imply the following conditions

$$\frac{1}{n}\log\Lambda_n^*(\Theta_i) \xrightarrow[n\to\infty]{\mathsf{P}_\theta-r-\text{quickly}} I_i(\theta) \quad \text{for all } \theta\in\Theta\setminus\Theta_i,\ i = 0, 1, \ldots, N, \tag{4}$$

where $I_i(\theta) = \inf_{\tilde{\theta}\in\Theta_i} I(\theta,\tilde{\theta})$ is assumed to be positive for all i. Let

$$J_i(\theta) = \min_{\substack{0\leq j\leq N\\ j\neq i}}[I_j(\theta)/c_{ji}] \quad \text{for} \quad \theta\in\Theta_i, \qquad J(\theta) = \max_{0\leq i\leq N} J_i(\theta) \quad \text{for} \quad \theta\in\mathrm{I}_{\mathrm{in}},$$

where $c_{ij} = \lim_{\alpha_{\max}\to 0}|\log\alpha_{ij}|/|\log\alpha_{\max}|$, $\alpha_{\max} = \max_{i,j}\alpha_{ij}$.

The following theorem establishes uniform asymptotic optimality of the MASLRT in the general non-iid case with respect to moments of the stopping time distribution. The proof is based on the technique developed by Tartakovsky [8] for multiple simple hypotheses.

Theorem 1 (MASLRT asymptotic optimality)**.** *Assume that r-quick convergence conditions* (2) *and* (4) *are satisfied. If the thresholds A_{ij} are so selected that $\sup_{\theta\in\Theta_i}\alpha_{ij}^*(\theta) \leq \alpha_{ij}$ and $\log A_{ij} \sim \log(1/\alpha_{ij})$, in particular $A_{ij} = 1/\alpha_{ij}$, then for $m \leq r$ as $\alpha_{\max}\to 0$*

$$\inf_{\delta\in\mathbf{C}(||\alpha_{ij}||)} \mathsf{E}_\theta T^m \sim \mathsf{E}_\theta[T^*]^m \sim \begin{cases}[|\log\alpha_{\max}|/J_i(\theta)]^m & \text{for all } \theta\in\Theta_i \text{ and } i = 0, 1, \ldots, N\\ [|\log\alpha_{\max}|/J(\theta)]^m & \text{for all } \theta\in\mathrm{I}_{\mathrm{in}}.\end{cases}$$

Consequently, the MASLRT minimizes asymptotically the moments of the sample size up to order r uniformly in $\theta \in \Theta$ in the class of tests $\mathbf{C}(||\alpha_{ij}||)$.

This theorem generalizes previous results of Pavlov [3] and Dragalin and Novikov [1] restricted to iid exponential families, and also provides alternative conditions in iid cases that can be often easily checked. Indeed, for a multidimensional exponential family, conditions (2) are satisfied for all $r > 0$ with $I(\theta, \tilde{\theta}) = \mathsf{E}_\theta \lambda_1(\theta, \tilde{\theta})$ being the Kullback–Leibler information numbers. Also, in many particular cases, conditions (4) also hold when $\hat{\theta}_n$ is the maximum likelihood estimator (MLE). For example, assume that $X_n \sim \mathcal{N}(\mu, \sigma^2)$, $n = 1, 2, \dots$ are iid normal random variables with unknown mean μ and unknown variance σ^2 and the hypotheses are "$H_0 : \mu = 0, \sigma^2 > 0$" and "$H_1 : \mu \geq \mu_1, \sigma^2 > 0$", where μ_1 is a given positive number. In this case, $N = 1$, $\theta = (\mu, \sigma^2)$ and the variance σ^2 is a nuisance parameter. It can be verified that if $(\hat{\mu}_n, \hat{\sigma}_n^2)$ is the MLE, $\hat{\mu}_n = \max\{0, n^{-1} \sum_{k=1}^n X_k\}$, $\hat{\sigma}_n^2 = n^{-1} \sum_{k=1}^n (X_k - \hat{\mu}_n)^2$, then conditions (4) hold for all $r > 0$ with $I_1(q) = \frac{1}{2} \log[1 + (q_1 - q)^2]$, $0 \leq q < q_1$ and $I_0(q) = \frac{1}{2} \log(1 + q^2)$, $q \geq 0$, where $q = \mu/\sigma$ and $q_1 = \mu_1/\sigma$. Therefore, by Theorem 1, the ASLRT minimizes (asymptotically) all positive moments of the sample size.

2. Sequential Changepoint Detection. Assume $X_1, X_2, \dots$ is a sequence of independent observations and $X_1, \dots, X_\nu$ have density $p_{\theta_0}(x)$ while at time ν something happens and $X_{\nu+1}, X_{\nu+2}, \dots$ have density $p_\theta(x)$, $\theta \in \Theta$, $\theta_0 \notin \Theta$. The pre-change parameter θ_0 is known, but the time of change $\nu \in \{0, 1, \dots\}$ and the post-change parameter θ are unknown. Let $W(\theta)$ be a weight (mixing prior distribution) and consider the following mixture-based Shiryaev–Roberts changepoint detection procedure

$$\mathsf{T}_{\mathsf{SR}}(A) = \inf \left\{ n \geq 1 : \int_\Theta R_n^\theta W(\mathrm{d}\theta) \geq A \right\}, \quad A > 0,$$

where $R_n^\theta = \sum_{k=1}^n \prod_{i=k}^n \frac{p_\theta(X_i)}{p_{\theta_0}(X_k)}$. We refer to this procedure as the WSR.

Let E_ν^θ denote expectation with respect to the probability measure P_ν^θ when the changepoint is ν and the post-change parameter is θ and let E_∞ denote expectation when there is no change. Let $\mathsf{ARL}(T) = \mathsf{E}_\infty T$ stand for the average run length (mean time) to false alarm. Let $\lambda_n^\theta = \sum_{k=1}^n \log \frac{p_\theta(X_k)}{p_{\theta_0}(X_k)}$ be the LLR and define the conditional expected Kullback–Leibler information $\mathcal{J}_\nu^\theta(T) := \mathsf{E}_\nu^\theta(\lambda_T^\theta - \lambda_\nu^\theta | T > \nu) = I_\theta \mathsf{E}_\nu^\theta(T - \nu | T > \nu)$, where $I_\theta = \mathsf{E}_0^\theta \lambda_1^\theta$. Then the maximal Kullback–Leibler information (over both ν and θ) is

$$\sup_{\theta \in \Theta} \sup_{\nu \geq 0} \mathcal{J}_\nu^\theta(T) = \sup_{\theta \in \Theta} \left[I_\theta \sup_{\nu \geq 0} \mathsf{E}_\nu^\theta(T - \nu | T > \nu) \right].$$

If p_θ belongs to the ℓ-dimensional exponential family, then the following two results can be established. First, the WSR procedure that starts off at zero ($R_0^\theta = 0$) is second order asymptotically optimal for any mixing distribution $W(\theta)$ with

continuous density in the class $\mathbf{C}(\gamma) = \{T : \mathsf{ARL}(T) \geq \gamma\}$:

$$\sup_{\theta\in\Theta,\nu\geq 0} \mathcal{J}_\nu^\theta(\mathsf{T}_{\mathsf{SR}}) = \inf_{T\in\mathbf{C}(\gamma)} \sup_{\theta\in\Theta,\nu\geq 0} \mathcal{J}_\nu^\theta(T) + O(1) \quad \text{as} \quad \gamma \to \infty,$$

where $O(1)$ is bounded as $\gamma \to \infty$. More importantly, if the WSR procedure starts off at a specially designed point and the mixing distribution $W = W^*$ is selected also in a special way depending on the average overshoot in the one-sided SPRT, then this specially designed WSR procedure $\mathsf{T}_{\mathsf{SR}}^*$ becomes third-order asymptotically optimal, i.e.,

$$\sup_{\theta\in\Theta,\nu\geq 0} \mathcal{J}_\nu^\theta(\mathsf{T}_{\mathsf{SR}}^*) = \inf_{T\in\mathbf{C}(\gamma)} \sup_{\theta\in\Theta,\nu\geq 0} \mathcal{J}_\nu^\theta(T) + o(1) \quad \text{as} \quad \gamma \to \infty,$$

where $o(1) \to 0$ as $\gamma \to \infty$.

The proofs are based on the works by Pollak [5, 4] and recent results of Fellouris and Tartakovsky [2].

Acknowledgements. This work was supported in part by the U.S. Air Force Office of Scientific Research under MURI grant FA9550-10-1-0569, by the U.S. Defense Threat Reduction Agency under grant HDTRA1-10-1-0086, by the U.S. Defense Advanced Research Projects Agency under grant W911NF-12-1-0034 and by the U.S. National Science Foundation under grants CCF-0830419 and EFRI-1025043 at the University of Southern California, Department of Mathematics.

References

[1] V. P. Dragalin and A. A. Novikov (1999). Adaptive sequential tests for composite hypotheses. *Surveys in Applied and Industrial Mathematics*, 6(2):387–398.

[2] G. Fellouris and A. G. Tartakovsky (2012). Nearly minimax one-sided mixture-based sequential tests. *Sequential Analysis*, 31(3) (in press).

[3] I. V. Pavlov (1990). Sequential procedure of testing composite hypotheses with applications to the Kiefer–Weiss problem. *Theory of Probability and its Applications*, 35(2):280–292.

[4] M. Pollak (1987). Average run lengths of an optimal method of detecting a change in distribution. *Annals of Statistics*, 15(2):749–779.

[5] M. Pollak (1978). Optimality and almost optimality of mixture stopping rules. *Annals of Statistics*, 6(4):910–916.

[6] H. Robbins and D. Siegmund (1970). A class of stopping rules for testing parameter hypotheses. In L. M. Le Cam, J. Neyman, and E. L. Scott, editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, June 21–July 18, 1970*, volume 4: Biology and Health, pages 37–41. University of California Press, Berkeley, CA, USA, 1972.

[7] H. Robbins and D. Siegmund (1974). The expected sample size of some tests of power one. *Annals of Statistics*, 2(3):415–436.

[8] A. G. Tartakovsky (1998). Asymptotic optimality of certain multihypothesis sequential tests: Non-i.i.d. case. *Statistical Inference for Stochastic Processes*, 1(3):265–295.

[9] Abraham Wald (1947). *Sequential Analysis*. John Wiley & Sons, Inc, New York, USA.