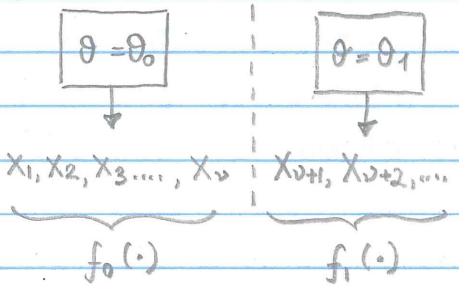
LECTURE 8

* Stochastic Models with Changes

Let us consider the stochastic process $\{X_k\}_k$ satisfying the condition:

- $\exists \nu \geq 0$: $X_1, X_2, \ldots, X_\nu$ are RVs that satisfy one distribution function $f_0(\cdot)$
  $X_{\nu+1}, X_{\nu+2}, \ldots$ are RVs that satisfy another distribution function $f_1(\cdot)$
  with $f_1(\cdot) \neq f_0(\cdot)$

Ex.:



The probability function of X_k depends on a parameter vector ϑ $\forall k$. The vector abruptly changes at time $\nu \geq 0$, i.e.:

$$f_0(x) \triangleq f_X(x \mid \vartheta = \vartheta_0)$$
$$f_1(x) \triangleq f_X(x \mid \vartheta = \vartheta_1)$$

Ex.: $X_1, X_2, \ldots, X_k, \ldots$ are observations of a latent RV (i.e., latent state) and the model (i.e., the probability function) of this RV changes over time, i.e.:

$$f_0(x) = f_{\vartheta_0}(x)$$
$$f_1(x) = f_{\vartheta_1}(x)$$

Ex.: A variation of the examples reported above can be obtained when the RVs $X_1, X_2, \ldots$ are not independent and the functions $f_0(\cdot), f_1(\cdot)$ are either conditional probability functions:

$$f_0 = f_0(X_k \mid X_{k-1}, X_{k-2}, \ldots, X_1)$$
$$f_1 = f_1(X_k \mid X_{k-1}, X_{k-2}, \ldots, X_1)$$

or they are joint probability functions.

In each and every case considered here, $\nu \geq 0$ is called CHANGE POINT

For a generic process $\{X_k\}$ with changepoint, the following two problems are typically of interest:

1) Offline Hypotheses Testing:

We have observations $(x_1, x_2, \dots, x_N)$ of $\{X_k\}$ collected up to $k=N>0$ and we must decide which one of the following two hypotheses is true:

H_0) $f_X(x_k | x_{k-1}, \dots, x_1) = f_0(x_k | x_{k-1}, \dots, x_1) \quad \forall k \in \{1, 2, 3, \dots, N\}$

H_1) $\exists \nu \geq 0 : f_X(x_k | x_{k-1}, \dots, x_1) = f_0(x_k | x_{k-1}, \dots, x_1) \quad k = 1, 2, \dots, \nu$

$f_X(x_k | x_{k-1}, \dots, x_1) = f_1(x_k | x_{k-1}, \dots, x_1) \quad k = \nu+1, \nu+2, \dots, N$

The problem here does NOT require to estimate ν, but just to determine if it exists

2) Sequential Hypothesis Testing:

We have an ongoing collection of observations of $\{X_k\}_k$ and, for any newly collected observation x_n, we must decide whether the sequence of observations $(x_1, x_2, \dots, x_n)$ is enough to be assigned to one of two (or more) possible classes, i.e., we must determine the value:

$$\hat{n} = \min \{ n : \exists i \geq 0 \text{ s.t. } (x_1, x_2, \dots, x_n) \sim f_i(\cdot) \}$$

In this case, we need a sequential test that, for every new n, determines if it is enough to make a terminal decision $\Rightarrow$ There will be a stopping time

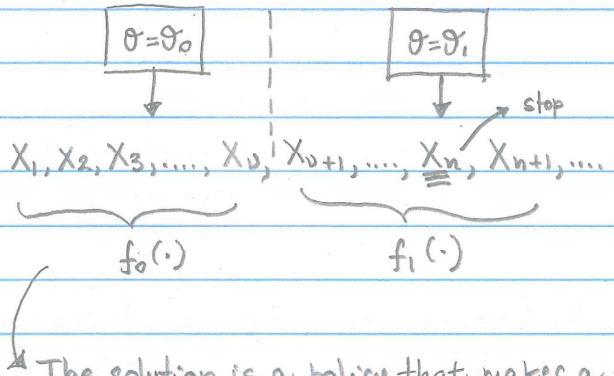
The solution is a policy that makes a tradeoff between average observation time and quality of the decision

The number of observations that is needed can be envisioned as a RV that must be estimated

3) Quickest Changepoint Detection:

We have an ongoing collection of observations of $\{X_k\}_k$ and, for any newly

collected observation x_n, we must decide whether a change point has occurred or not ⇒ The stopping time is now the time when we have enough observations to conclude that a change has occurred



Regardless of the value of ν, the stopping time is when we have enough data to decide in favour of a change ⇒ $n \geq \nu$, i.e., a delay is intrinsically related to the problem

The solution is a policy that makes a tradeoff between detection delay and possibility of a FALSE POSITIVE (i.e., $n < \nu$ or $n \in \mathbb{Z}_0^+$ and $\nu = \infty$)

Note: Problem 2 and 3 are solved by choosing a decision policy ⇒ A decision policy is a pair $\delta = (T, d)$ defined by:
$T \triangleq$ stopping time
$d \triangleq$ terminal decision

In Problem 3, we can have two scenarios: (i) there are only N=2 classes for the observations $(x_1, x_2, \ldots)$ ⇒ "d" is trivial and follows the choice of T - or the scenario (ii) there are N>2 classes ⇒ "d" is the decision about which post-change hypothesis is satisfied (i.e., which one of the N-1 classes the observations $x_{\nu+1}, x_{\nu+2}, \ldots$ belong to)

Note: The decision policy δ is supposed to be built (or updated) at every new observation (⇒ online implementation) and the management of the tradeoff can be eventually done by minimizing a cost function ⇒ An optimization problem can be formulated

We will specifically focus on algorithms that solve the Quickest Change-point Detection problem.

* Change point Models

A change point model describes the probabilistic structure of the process $\{X_k\}_k$ and of the change point ν. Numerous options are possible:

- ν — R.V (with a distribution that is know a priori)
  - unknown but non-random number
- X_k — i.i.d
  - independent but nonidentically distributed
  - dependent observations

Let us consider the history $\mathcal{H}_n \triangleq (X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1})$ and let us denote with $f_{n,0}(x)$ and $f_{n,1}(x)$ the pdf of X_n under condition $\vartheta = \vartheta_0$ (i.e., before change point) and $\vartheta = \vartheta_1$ (i.e., after change point), respectively. We have:

$$f_\nu(X_1, X_2, X_3, \dots, X_n) = \prod_{k=1}^{\nu} f_{k,0}(X_k \mid \mathcal{H}_k) \prod_{k=\nu+1}^{n} f_{k,1}(X_k \mid \mathcal{H}_k)$$

joint probability function when the change point is ν — if $0 \le \nu \le n$

In general, if X_k are dependent on the change point (e.g., X_k does not depend on $\mathcal{H}_k$ but only on $(X_{\nu+1}, X_{\nu+2}, \dots, X_{k-1})$ after the change point), we can assume that $f_{k,1}(\cdot)$ is affected by ν and write:

$$f_\nu(X_1, X_2, \dots, X_n) = \prod_{k=1}^{\nu} f_{k,0}(X_k \mid \mathcal{H}_k) \prod_{k=\nu+1}^{n} f_{k,1}^{(\nu)}(X_k \mid \mathcal{H}_{\nu+1}^{k}) \qquad (*)$$

Note: X_k - i.i.d. $\Rightarrow f_{k,0}(X_k \mid \mathcal{H}_k) = f_0(X_k)$ $\qquad$ $\mathcal{H}_{\nu+1}^{k} \triangleq (X_{\nu+1}, \dots, X_{k-1})$

$f_{k,1}^{(\nu)}(X_k \mid \mathcal{H}_k) = f_1(X_k)$

Also note that, in realistic scenarios, $f_{k,0}(\cdot)$ are known (or they belong to a known class) while $f_{k,1}(\cdot)$ may be unknown and their estimation may be unnecessary in order to choose $\delta \Rightarrow$ The estimation of the post-change distribution functions is necessary if there are $N-1 > 1$ possible classes of post-change distribution functions (CHANGE DETECTION AND ISOLATION PROBLEM)

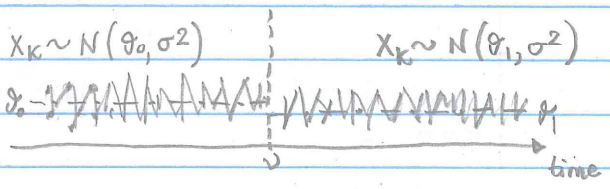
Finally, it is possible that, in a time series, there are multiple changepoints ⇒ We typically formulate as many changepoint detection problems as the number of changepoints ⇒ A reset of δ is assumed at every new detection

Let us consider the changepoint ν. If ν is a RV, its probability function may depend on the observations $X_1, X_2, X_3, \dots$ ⇒ We introduce:

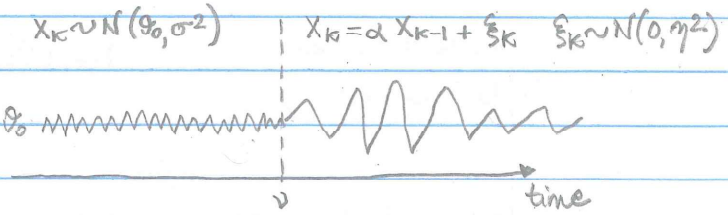
$$q \triangleq P(\nu < 0) \qquad (**)$$
$$\pi_k \triangleq P(\nu < k \mid X_k, \mathcal{H}_k)$$

Note that a change point model is defined when a model is given for (*) (at least, for $f_{k,0}(\cdot)$) and (**). By looking at the models for (*), two classes of changes have usually practical interest:

Additive Change ↑
- change in the mean value of $\{X_k\}_k$



Non additive Change ↑
- change in the mechanism of generation of $\{X_k\}_k$



Additive changes do not affect the probabilistic structure of the process $\{X_k\}$ (i.e., if X_k are observations of a dynamical system, the system dynamics is not affected). Nonadditive changes, instead, are changes that affect the variance, correlations or other nonadditive features of $\{X_k\}_k$ ⇒ The system dynamics is affected

By looking at the models for (**), we can quantify the possibility (a.k.a. "risk") of a false positive via: $\underset{T}{E}(T \mid \nu = \infty) \triangleq$ mean time to false positive

Analogously, we can measure the average detection delay via:

$$\underset{T}{E}\left(T-\nu \mid T>\nu\right) \qquad \nu=0,1,2,\ldots$$

* Algorithms for Changepoint Detection

Let us assume, for sake of simplicity, the following conditions:

- $\{X_k\}_k$ is made of independent RVs with a probability density $f_X(\cdot)$ conditioned on a scalar $\vartheta \Rightarrow$ Let us define: $f_\vartheta(x) \triangleq f_X(x|\vartheta)$

- $\left.\begin{array}{l}\vartheta=\vartheta_0 \text{ for } k\leq\nu \\ \vartheta=\vartheta_1 \text{ for } k>\nu\end{array}\right\}$ with $\nu\geq 0$ changepoint. ϑ_0 is known, ϑ_1 may be unknown $\Rightarrow$ Let us define: $f_i(x) \triangleq f_{\vartheta_i}(x)$

Let us also the following functions:

Log-likelihood Ratio (LLR): $$LLR(x) \triangleq \ln \frac{f_1(x)}{f_0(x)} \qquad \forall x$$

Kullback-Leibler Divergence (KL): $$KL(\vartheta_1,\vartheta_0) \triangleq \underset{\vartheta_1}{E}\left[\ln \frac{f_1(X)}{f_0(X)}\right] = \underset{\vartheta_1}{E}\left[LLR\right]$$

A typical detection algorithm requires a function that tracks the probabilistic structure of $\{X_k\}_k$ as the number of observations increases and a threshold to decide when the change has occurred $\Rightarrow$ A simple solution is:

- Function: $S_i^j \triangleq \sum_{k=i}^{j} LLR(x_k) \qquad \forall i,j$

Chart-based Algorithms

- Hypotheses: $H_0)\ \vartheta=\vartheta_0 \qquad H_1)\ \vartheta=\vartheta_1$

- Decision: $d=\begin{cases}0 & \text{if } S_1^N < h \Rightarrow H_0 \text{ is chosen} \\ 1 & \text{if } S_1^N \geq h \Rightarrow H_1 \text{ is chosen}\end{cases}$

with h - threshold to be chosen and $N\geq 1$ fixed and known

In this case, we need N observations to obtain one value of $S_1^N \Rightarrow$ The decision policy is updated every N samples and the stopping time is given by the rule:

$$T = N \cdot \min\{k : d_k = 1\}$$

where d_k is the decision taken by using the k-th batch of N observations

Ex.: $X_k \sim N(\vartheta, \sigma^2)$ - We have:

$$f_\vartheta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\vartheta)^2}{2\sigma^2}} \Rightarrow LLR(x) = -\frac{(x-\vartheta_1)^2}{2\sigma^2} + \frac{(x-\vartheta_0)^2}{2\sigma^2} = \frac{\vartheta_1-\vartheta_0}{\sigma^2}\left(x - \frac{\vartheta_0+\vartheta_1}{2}\right)$$

Hence: $S_1^N = \frac{\vartheta_1-\vartheta_0}{\sigma^2}\sum_{k=1}^{N}\left(x_k - \frac{\vartheta_0+\vartheta_1}{2}\right) = \frac{\vartheta_1-\vartheta_0}{\sigma^2}\left[\sum_{k=1}^{N} x_k - \frac{N}{2}(\vartheta_0+\vartheta_1)\right]$

At the generic batch K, we have:

$$S_1^N(K) = S_{N(K-1)+1}^{NK}$$

In this example, we need that both ϑ_0 and ϑ_1 are known. The decision function is typically formulated as:

$$\left.\begin{array}{l} b \triangleq \frac{\vartheta_1-\vartheta_0}{\sigma} \\ \gamma \triangleq \vartheta_1-\vartheta_0 \end{array}\right\} \Rightarrow S_1^N = \frac{b}{\sigma} N\left(\bar{x} - \vartheta_0 - \frac{\gamma}{2}\right) \qquad \bar{x} \triangleq \frac{1}{N}\sum_{k=1}^{N} x_k \text{ - sample mean}$$

The threshold is empirically chosen as $h \triangleq \lambda b\sqrt{N} - \frac{N}{2}\left(\frac{\vartheta_1-\vartheta_0}{\sigma}\right)^2$ with λ to be set. In this way:

$$S_1^N(k) \geq h \Leftrightarrow \bar{x}_K \geq \vartheta_0 + \lambda\frac{\sigma}{\sqrt{N}}$$

Shewhart Control Chart

with $\bar{x}_K \triangleq \frac{1}{N}\sum_{i=1}^{N} x_{N(K-1)+i}$ and the assumption: $\vartheta_1 > \vartheta_0$

□

An alternative to the previous algorithm can be conceived to avoid the need for batches of N observations per iteration and to give higher relevance to the latest observations:

$$g_k \triangleq \sum_{i=0}^{\infty} \gamma_i \, LLR(x_{k-i}) \quad \Leftarrow \text{New decision function}$$

$$\gamma_i \triangleq \alpha(1-\alpha)^i \qquad 0<\alpha\leq 1 \quad \Leftarrow \text{Exponential weights}$$

In this case, we have: $g_k = g_{k-1}(1-\alpha) + \alpha \, LLR(x_k)$

Geometric Moving Average Chart

$$T \triangleq \min\{k : g_k \geq h\}$$

Ex: $X_k \sim N(\mu, \theta^2) \Rightarrow$ We can write:

$$LLR(x_k) = \ln\frac{\theta_0}{\theta_1} + \left(\frac{1}{\theta_0^2} - \frac{1}{\theta_1^2}\right)\frac{(x_k-\mu)^2}{2}$$

$$\Rightarrow g_k = \ln\frac{\theta_0}{\theta_1}\sum_{i=0}^{\infty}\gamma_i + \left(\frac{1}{\theta_0^2} - \frac{1}{\theta_1^2}\right)\sum_{i=0}^{\infty}\gamma_i \frac{(x_{k-i}-\mu)^2}{2}$$

$$= (1-\alpha)g_{k-1} + \alpha \, LLR(x_k)$$

If we define: $\tilde{g}_k \triangleq \frac{2\theta_0^2\theta_1^2}{\theta_1^2-\theta_0^2} g_k - \frac{2\theta_0^2\theta_1^2}{\theta_1^2-\theta_0^2}\ln\frac{\theta_0}{\theta_1}$

we have: $\tilde{g}_k = \frac{2\theta_0^2\theta_1^2}{\theta_1^2-\theta_0^2}\left[(1-\alpha)g_{k-1} + \alpha \, LLR(x_k) - \ln\frac{\theta_0}{\theta_1}\right]$

$$= \frac{2\theta_0^2\theta_1^2}{\theta_1^2-\theta_0^2}\left[(1-\alpha)\left(g_{k-1} - \ln\frac{\theta_0}{\theta_1}\right)\right] + \alpha(x_k-\mu)^2$$

$$= (1-\alpha)\tilde{g}_{k-1} + \alpha(x_k-\mu)^2$$

□

Note that the Geometric Moving Average consists of running a infinite-memory filter on the values of LLR $\Rightarrow$ Variants can be defined with finite-memory filters (i.e., finite number of weights γ_i)

By following the same approach, one can use the discrete derivative

$$\nabla g_k \triangleq g_k - g_{k-1}$$

as the decision function. In this case, we are typically interested in counting how many times $|\nabla g_k|$ crosses a given threshold, i.e., we define:

$$I_{\nabla g_k} \triangleq \begin{cases} 1 & |\nabla g_k| \geq h \\ 0 & |\nabla g_k| < h \end{cases}$$ - Then, we consider the decision rule:

$$d \triangleq \begin{cases} 1 & \sum_{i=0}^{N-1} I_{\nabla g_{k-i}} \geq \eta \\ 0 & \text{otherwise} \end{cases} \Rightarrow \text{Hence, } T \triangleq \min\left\{ k : \sum_{i=0}^{N-1} I_{\nabla g_{k-i}} \geq \eta \right\}$$

In these rules, h and η are thresholds to be assigned and N defines the observation window that is monitored to make a decision. □

These algorithms, though, rely on using fixed thresholds ⇒ No adaptation to the dynamics of the process. Also, a decision is taken the very first time a quantity derived from LLR is above threshold ⇒ There may be sensitivity to noise. An alternative solution that copes with these issues is:

CUSUM Algorithm

- Function: $g_k \triangleq S_1^k - m_k$, $m_k \triangleq \min_{1 \leq j \leq k} S_1^j$

- Decision: $d = \begin{cases} 0 & \text{if } g_k < h \\ 1 & \text{if } g_k \geq h \end{cases}$ $h \triangleq$ threshold to be chosen

Note: $g_k \geq h \iff S_1^k \geq m_k + h$ ⇒ The threshold is adaptive and depends on the evolution of the cumulative sum S_1^j over time $j \leq k$

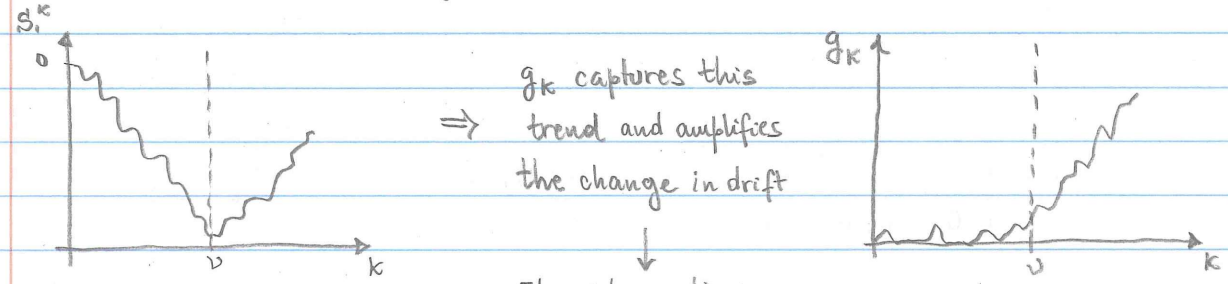
Also, note: $S_1^k = \min_{1 \leq j \leq k} S_1^j \Rightarrow g_k = 0$

$S_1^k > \min_{1 \leq j \leq k} S_1^j \Rightarrow m_k = m_{k-1}$ and $g_k = g_{k-1} + LLR(x_k)$

Therefore, we can formulate the decision function in a recursive way:

$$g_k = \sup\left(0,\ g_{k-1} + LLR(x_k)\right) \quad \leftarrow \text{It is a RANDOM WALK}$$

From a graphical standpoint, one may note: $f_0(x) > f_1(x) \Rightarrow LLR(x) < 0 \Rightarrow$ The cumulative sum S_1^k has a negative drift for $k \leq \nu$ and eventually changes as $k > \nu$



$\Rightarrow$ g_k captures this trend and amplifies the change in drift

It resets anytime the value of S_k decreases

Note: the negative drift of S_1^k for $k \leq \nu$ can be estimated by computing the expected value of LLR:

$$\underset{x_k}{E}\left(LLR(x_k)|\nu\right) = \int_{-\infty}^{\infty} f_0(x)\, LLR(x)\, dx = \underset{\vartheta_0}{E}\left[LLR\right] = -KL(\vartheta_0, \vartheta_1) < 0 \qquad \nu \geq k$$

$$\underset{x_k}{E}\left(LLR(x_k)|\nu\right) = \int_{-\infty}^{+\infty} f_1(x)\, LLR(x)\, dx = KL(\vartheta_1, \vartheta_0) > 0 \qquad \nu < k$$

An interesting generalization can be obtained in the Gaussian case with additive change:

$$\left.\begin{array}{ll} X_k \sim N(\vartheta_0, \sigma) & k \leq \nu \\ X_k \sim N(\vartheta_1, \sigma) & k > \nu \end{array}\right\} \Rightarrow LLR(x) = \frac{b}{\sigma}\left(x - \vartheta_0 - \frac{\gamma}{2}\right)$$

$$\Rightarrow g_k = \sup\left(0,\ g_{k-1} + \frac{b}{\sigma}\left(x_k - \vartheta_0 - \gamma/2\right)\right)$$

Now, let us suppose that ϑ_1 can assume either the value: $\vartheta_1^+ \triangleq \vartheta_0 + \delta$ or the value: $\vartheta_1^- \triangleq \vartheta_0 - \delta$ with $\delta > 0$ fixed and known (e.g., $\delta = \sigma$) $\Rightarrow LLR(x_k)$ can be one of these two values:

$$LLR(x_k)^+ = \frac{\delta}{\sigma}\left(x_k - \vartheta_0 - \frac{\delta}{2}\right) \overset{\delta=\sigma}{=} x_k - \vartheta_0 - \delta/2 \quad \text{if } \vartheta_1 = \vartheta_1^+$$

$$LLR(x_k)^- = \frac{\delta}{\sigma}\left(-x_k + \vartheta_0 - \frac{\delta}{2}\right) \overset{\delta=\sigma}{=} -x_k + \vartheta_0 - \delta/2 \quad \text{if } \vartheta_1 = \vartheta_1^-$$

Therefore, we can define two decision functions:

$$\left.\begin{aligned} g_k^+ &= \sup\left(0,\ g_{k-1}^+ + LLR(x_k)^+\right) \\ g_k^- &= \sup\left(0,\ g_{k-1}^- + LLR(x_k)^-\right) \end{aligned}\right\} \Rightarrow \text{The decision rule becomes:}$$

$$d \triangleq \begin{cases} 1 & g_k^+ \geq h \ \cup\ g_k^- \geq h \\ 0 & \text{otherwise.} \end{cases}$$

Note that such decision rule corresponds to introduce two straight line thresholds for $\sum_{i=1}^{k} x_i$ (a.k.a. "V-mask") and detect a change when one of them is hit:

$$g_k^+ = \sup\left(0,\ g_{k-1}^+ + x_k - \vartheta_0 - \frac{\delta}{2}\right) = \max_{1\leq j\leq k} \sum_{i=j}^{k}\left(x_i - \vartheta_0 - \frac{\delta}{2}\right) \Rightarrow$$

$$g_k^+ \geq h \overset{\text{EQUIV}}{\Longleftrightarrow} \max_{1\leq j\leq k}\left[\sum_{i=j}^{k} x_i - (k-j)(\vartheta_0 - \delta/2)\right] \geq h \overset{\text{EQUIV}}{\Longleftrightarrow}$$

Denoted with j^* the arg of the max, we have: $\sum_{i=j^*}^{k} x_i \geq h + (k-j^*)(\vartheta_0 - \delta/2)$ ← It's a line □

A different class of algorithms can be developed under the assumptions:

Bayesian Algorithms

- ν is a RV
- The distribution of ν is known A PRIORI

→ The Bayes' Theorem can be invoked

In particular, we can introduce:

- Probability of false positive: $P(T \leq \nu) = \sum_{k=1}^{\infty} P(T \leq \nu \mid \nu = k) P(\nu = k) \triangleq P^{\pi}(T \leq \nu)$

- Average delay to a detection: $\underset{T}{E}(T - \nu \mid T > \nu) = \int_0^{+\infty} (T-\nu) \frac{f(T-\nu, T>\nu)}{P(T>\nu)} dT$

$$= \int_0^{+\infty} (T-\nu) \frac{\sum_k f(T-\nu, T>\nu \mid \nu=k) P(\nu=k)}{\sum_k P(T>\nu \mid \nu=k) P(\nu=k)} dT = \frac{1}{P^{\pi}(T>\nu)} \sum_k P(\nu=k) \int_0^{+\infty} (T-\nu) f(T-\nu, T>\nu \mid \nu=k) dT$$

In this formula :

$k > T \Rightarrow T - \nu < 0$ and $f(T-\nu, T>\nu \mid \nu = k) = 0$ - Therefore, we can define:

$(T-\nu)^+ \triangleq \max(0, T-\nu)$ and write:

$$\int_0^{+\infty} (T-\nu) f(T-\nu, T>\nu \mid \nu=k) dT = \underset{T}{E}\left((T-\nu)^+ \mid \nu=k\right)$$

Hence we have: $\underset{T}{E}(T-\nu \mid T>\nu) = \frac{1}{P^\pi(T>\nu)} \underbrace{\sum_k \underset{T}{E}\left((T-\nu)^+ \mid \nu=k\right) P(\nu=k)}_{E^\pi((T-\nu)^+)} = \frac{E^\pi((T-\nu)^+)}{P^\pi(T>\nu)}$

Let us call : $\text{PFA}(T) \triangleq P^\pi(T \le \nu)$

$\text{ADD}(T) \triangleq E^\pi((T-\nu)^+) / P^\pi(T>\nu)$

} - We can now derive the decision

policy by solving an optimization problem that explicitly minimizes the tradeoff between PFA(T) and ADD(T) $\Rightarrow$ We can represent this tradeoff in a cost function to be minimized.

- What is a good model for $P(\nu=k)$?

In general, if no prior knowledge is given about ν, it is assumed that:

$$\nu \sim \text{Geom}(\rho)$$

i.e., $P(\nu=k) = \rho(1-\rho)^{k-1} \quad \forall k>0$

This assumption simplifies the calculations:

$$P(\vartheta=\vartheta_1 \text{ at time } k \mid \vartheta=\vartheta_0 \text{ at time } k-1) = P(\nu=k-1 \mid \nu \ge k-1) = \frac{P(\nu=k-1)}{P(\nu \ge k-1)}$$

$$= \frac{\rho(1-\rho)^{k-2}}{\rho(1-\rho)^{k-2}\sum_{m=0}^{\infty}(1-\rho)^m} = \rho \quad \forall k>0$$

$$P(\vartheta=\vartheta_0 \text{ at time } k \mid \vartheta=\vartheta_0 \text{ at time } k-1) = P(\nu \geq k \mid \nu \geq k-1) = 1-p \quad \forall k \geq 1$$

And similarly:

$$P(\vartheta=\vartheta_1 \text{ at time } k \mid \vartheta=\vartheta_1 \text{ at time } k-1) = P(\nu < k \mid \nu < k-1) = \frac{P(\nu<k-1)}{P(\nu<k-1)} = 1$$

$$P(\vartheta=\vartheta_0 \text{ at time } k \mid \vartheta=\vartheta_1 \text{ at time } k-1) = 0$$

By using the a priori distribution function $P(\nu=k)$ we can compute the a posteriori probability of a change in (**):

$$\pi_k \triangleq P(\nu<k \mid X_k, \mathcal{H}_k) = \frac{P(X_k \mid \mathcal{H}_k, \nu<k)\, P(\nu<k \mid \mathcal{H}_k)}{P(X_k \mid \mathcal{H}_k)}$$

$P(X_k \mid \mathcal{H}_k, \nu<k) = f_1(X_k)$ ← distribution function when $\vartheta=\vartheta_1$

$$P(X_k \mid \mathcal{H}_k) = P(X_k \mid \mathcal{H}_k, \nu<k)\, P(\nu<k \mid \mathcal{H}_k) + P(X_k \mid \mathcal{H}_k, \nu \geq k)\, P(\nu \geq k \mid \mathcal{H}_k)$$

$$= f_1(X_k)\, P(\nu<k \mid \mathcal{H}_k) + f_0(X_k)\, P(\nu \geq k \mid \mathcal{H}_k)$$

Hence, by denoting: $L(x) \triangleq f_1(x)/f_0(x)$ - likelihood ratio, we can write:

$$\pi_k = \frac{L(X_k)\, P(\nu<k \mid \mathcal{H}_k)}{L(X_k)\, P(\nu<k \mid \mathcal{H}_k) + P(\nu \geq k \mid \mathcal{H}_k)}$$

We can further develop the relationship:

$$P(\nu<k \mid \mathcal{H}_k) = P(\nu<k-1 \mid X_{k-1}, \mathcal{H}_{k-1}) + P(\nu=k-1 \mid \nu \geq k-1, \mathcal{H}_k)\, P(\nu \geq k-1 \mid \mathcal{H}_k)$$

$$= \pi_{k-1} + p(1-\pi_{k-1})$$

$$P(\nu \geq k \mid \mathcal{H}_k) = P(\nu \geq k \mid \nu \geq k-1, \mathcal{H}_k)\, P(\nu \geq k-1 \mid \mathcal{H}_k) = (1-p)(1-\pi_{k-1})$$

where we have used the facts:

$$P(\nu = k-1 \mid \nu \geq k-1, \mathcal{H}_k) = P(\vartheta = \vartheta_1 \text{ at time } k \mid \vartheta = \vartheta_0 \text{ at time } k-1) = \rho$$

$\mathcal{H}_k$ is the sequence up to $k-1$ and does not provide any further information

Prior distribution is geometric

$$P(\nu \geq k-1 \mid \mathcal{H}_k) = 1 - P(\nu < k-1 \mid X_{k-1}, \mathcal{H}_{k-1}) = 1 - \pi_{k-1}$$

$$P(\nu \geq k \mid \nu \geq k-1, \mathcal{H}_k) = P(\vartheta = \vartheta_0 \text{ at time } k \mid \vartheta = \vartheta_0 \text{ at time } k-1) = 1-\rho$$

Therefore, a recursive formulation of π_k in (**) is achieved:

$$\pi_k = \frac{L(x_k)\left[\pi_{k-1} + \rho(1-\pi_{k-1})\right]}{(1-\rho)(1-\pi_{k-1}) + L(x_k)\left[\pi_{k-1} + \rho(1-\pi_{k-1})\right]} \qquad (***)$$

which can be further simplified by defining: $\omega_k \triangleq \frac{\pi_k}{1-\pi_k}$:

$$\omega_k = \frac{L(x_k)}{1-\rho}(\omega_{k-1} + \rho) \Rightarrow \text{or, in log-scale: } g_k = \log\left(e^{g_{k-1}} + \rho\right) + LLR(x_k) - \log(1-\rho)$$

$g_k \triangleq \log \omega_k$

In summary:

Bayesian Approach $\Rightarrow$ We can give a measure for false positives (PFA) and delay in detection (ADD) $\Rightarrow$ We can build a cost function and solve an optimization problem to get the decision policy δ

$\nu \sim \text{Geom}(\rho) \Rightarrow$ Calculations simplify and a recursive formulation for π_k is given $\Rightarrow$ The function g_k is not a random walk but it has a specific meaning related to the latent state (it depends on π_k)

References:

Basseville, M. and Nikiforov, I. V. (1993) "Detection of Abrupt Changes: Theory and Applications", Prentice Hall Ed., Englewood Cliffs, NJ

↑

I referred to ch. 1-2 - A copy of the book is on HuskyCT