CS/CNS/EE/IDS 165: Foundations of Machine Learning and Statistical Inference

Sequential Detection

http://tensorlab.cms.caltech.edu/users/anima/cms165-2020.html

Anima Anandkumar
Computing and Mathematical Sciences
California Institute of Technology, Pasadena CA 91125
anima@caltech.edu

Outline

Concepts

- Motivation.
- Sequential Probability Ratio Test (SPRT)
- Optimality of SPRT

References

1. H.V. Poor, An Introduction to Signal Detection and Estimation, 2nd Ed., Springer-Verlag, 1994, Chapter III.
2. E.L. Lehmann, Testing Statistical Hypotheses, John Wiley & Sons, Inc., 1959, Chapter 3.
3. T. S. Ferguson, Mathematical Statistics: A Decision Theoretic Approach, Academic Press, 1967, Chapter 7.

Motivation

Example Consider the n-sample simple binary hypotheses

$$\mathcal{H}_0 : Y_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1) \quad \text{vs.} \quad \mathcal{H}_1 : Y_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(1,1), \quad i = 1, \cdots, n$$

What is the minimum n such that the optimal detector has size less then α and power greater than β?

The optimal detector tests the likelihood ratio against a threshold, and the log likelihood ratio of sample size n is

$$\ln L(\mathbf{Y}_n) = \sum_{i=1}^{n} Y_i - \frac{n}{2} \sim \begin{cases} \mathcal{N}(-\frac{n}{2}, n) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(\frac{n}{2}, n) & \text{under } \mathcal{H}_1 \end{cases}$$

We need to determine the threshold τ and sample size n to satisfy the size and power requirements.

$$\alpha = Q(\frac{\frac{n}{2} + \tau}{\sqrt{n}}) \quad \beta = Q(\frac{-n/2 + \tau}{\sqrt{n}})$$

The minimum sample size for such a fixed sample size (FSS) detector is

$$n_{\text{FSS}} \geq [Q^{-1}(\alpha) - Q^{-1}(\beta)]^2$$

When $\alpha = 1 - \beta = Q(4) \approx 3.16 \times 10^{-5}$, $n_{\text{FSS}} = 64$.

What if we are lucky...

Suppose that the first sample we receive is $Y = 20$? It seems that it is far more likely that this sample is from $\mathcal{N}(1,1)$ than from $\mathcal{N}(0,1)$. Perhaps a decision can already be rendered.

Another example is $\mathcal{H}_0 : \mathcal{U}(0,1)$ vs. $\mathcal{H}_1 : \mathcal{U}(0.5, 1.5)$. It is obvious that if we receive a sample in $(0, 0.5) \bigcup (1, 1.5)$, we can make decision directly without error.

Sequential Detection

Defining a Sequential Detector

Consider simple binary hypotheses $\mathcal{H}_0$ vs. $\mathcal{H}_1$

$$\mathcal{H}_i : Y_k \overset{\text{i.i.d.}}{\sim} f(y|\theta_i), k = 1, 2, \cdots$$

We will use the notation $\mathbf{Y}_n \overset{\Delta}{=} [Y_1, \cdots, Y_n]'$.

A sequential detector is defined by the sequence of stopping rules $\boldsymbol{\phi} = (\phi_i)$ and its corresponding decision rules $\boldsymbol{\delta} = (\delta_i)$.

- Stopping Rule The stopping rule is a binary function $\phi_n : \mathcal{R}^n \to \{0, 1\}$ of $\mathbf{Y}_n$. For $\mathbf{Y}_n = \mathbf{y}_n$, data collection stops if $\phi_n(\mathbf{y}_n) = 1$ and continues otherwise. For a sequence of random observations $\{\mathbf{Y}_n\}$, The stop time is a random variable

$$N(\boldsymbol{\phi}) = \min\{k : \phi_k(\mathbf{Y}_k) = 1\}.$$

- Terminal Decision Rule The terminal decision rule is given by

$$\delta(\mathbf{y}_n) = \Pr[D = 1|\mathbf{Y}_n = \mathbf{y}_n].$$

We will denote a sequential detector by the pair $(\boldsymbol{\phi}, \boldsymbol{\delta})$ where $\boldsymbol{\phi} = [\phi_i]$ and $\boldsymbol{\delta} = [\delta_i]$.

Sequential Probability Ratio Test

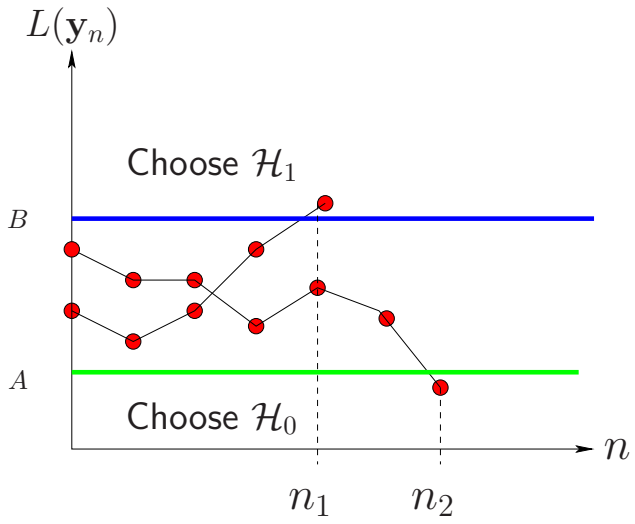
The Sequential Probability Ratio Test (SPRT)

The sequential probability ratio test (SPRT) denoted by SPRT(A,B) is based on the likelihood ratio

$$L(\mathbf{y}_n) = \frac{f(\mathbf{y}_n|\theta_1)}{f(\mathbf{y}_n|\theta_0)}$$

$$\phi_n(\mathbf{y}) = \begin{cases} 0 & L(\mathbf{y}_n) \in (A, B) \\ 1 & o.w. \end{cases}$$

$$\delta_n(\mathbf{y}) = \begin{cases} 0 & L(\mathbf{y}_n) \leq A \\ 1 & L(\mathbf{y}_n) \geq B \end{cases}$$



The Wald-Wolfowitz Theorem

The SPRT(A,B) detector $(\boldsymbol{\phi}_*, \boldsymbol{\delta}_*)$ has the minimum average stop time among all detectors with size no larger and power no less than those of $(\boldsymbol{\phi}_*, \boldsymbol{\delta}_*)$. Specifically,

$$\begin{matrix} P_{\mathsf{F}}(\delta) \leq P_{\mathsf{F}}(\delta_*) \\ P_{\mathsf{D}}(\delta) \geq P_{\mathsf{D}}(\delta_*) \end{matrix} \rightarrow \mathbb{E}_{\theta_i}(N(\boldsymbol{\phi}_*)) \leq \mathbb{E}_{\theta_i}(N(\boldsymbol{\phi})) \quad i = 0, 1.$$

Remark We need to determine A and B from power β and size α, which turns out to be difficult. Intuitively, larger the gap $B - A$, more reliable the decision and longer the stop time.

Determining Thresholds

Bounds on False Alarm and Miss Detection

Given size α and power β, consider the event that SPRT(A,B) has a false alarm. Let

$$\Gamma_1^{(n)} = \{\mathbf{y}_n; L(\mathbf{y}_n) \geq B, L(\mathbf{y}_k) \in (A,B), \forall k < n\}$$

We have

$$\begin{aligned} \alpha &= \sum_{n=1}^{\infty} \int_{\Gamma_1^{(n)}} f(\mathbf{y}_n|\theta_0) d\mathbf{y}_n \\ &\leq \frac{1}{B} \sum_{n=1}^{\infty} \int_{\Gamma_1^{(n)}} f(\mathbf{y}_n|\theta_1) d\mathbf{y}_n = \frac{\beta}{B} \end{aligned}$$

Given size α and power β, we should choose $B \leq \frac{\beta}{\alpha} \stackrel{\Delta}{=} B'$

Similarly, we should choose $A \geq \frac{1-\beta}{1-\alpha} \stackrel{\Delta}{=} A'$ because

$$\begin{aligned} 1-\beta &= \sum_{n=1}^{\infty} \int_{\Gamma_0^{(n)}} f(\mathbf{y}_n|\theta_1) d\mathbf{y}_n \\ &\leq A \sum_{n=1}^{\infty} \int_{\Gamma_0^{(n)}} f(\mathbf{y}_n|\theta_0) d\mathbf{y}_n = A(1-\alpha) \end{aligned}$$

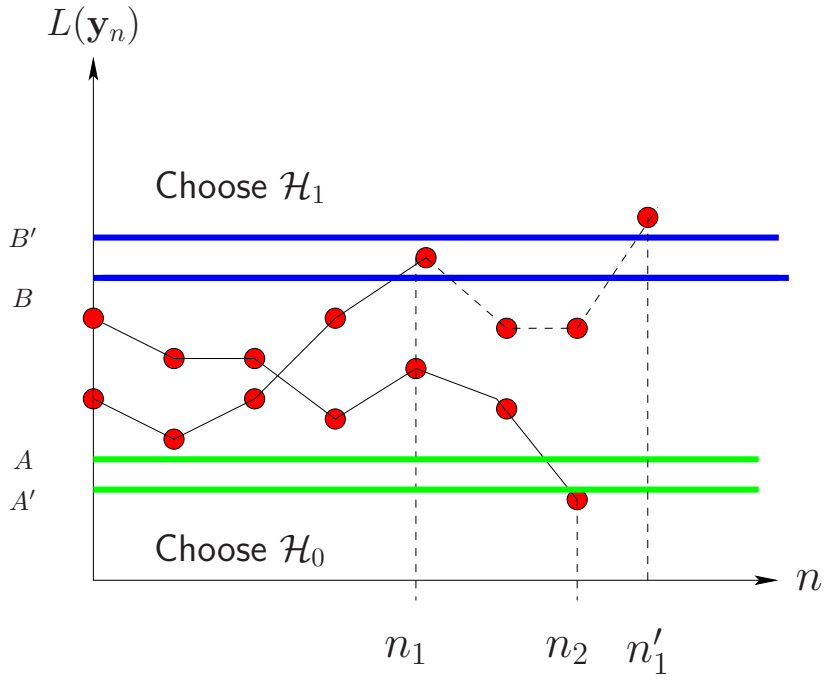
where $\Gamma_0^{(n)}$ is the set of $\mathbf{y}_n$ that miss detection occurs.

Wald's Approximations

Given size α and power β, the optimal SPRT(A,B) can be approximated by SPRT(A',B') with

$$A' = \frac{1-\beta}{1-\alpha}, \quad B' = \frac{\beta}{\alpha}$$

Justifications of Wald's Approximations



- Since $(A, B) \subseteq (A', B')$, more samples are required, and SPRT(A',B') has similar power and size as that of SPRT(A,B):

$$\begin{cases} \frac{\beta}{\alpha} & \leq & \frac{\beta'}{\alpha'} \\ \frac{1-\beta'}{1-\alpha'} & \leq & \frac{1-\beta}{1-\alpha} \end{cases} \rightarrow \begin{cases} \alpha' & \leq & \alpha\frac{\beta'}{\beta} \\ 1-\beta' & \leq & (1-\beta)\frac{1-\alpha'}{1-\alpha} \\ \alpha'+1-\beta' & \leq & \alpha+1-\beta \end{cases}$$

Average Sample Size

Lemma (The Wald's Equation)

Let Z_i be independent and identically distributed with $\mathbb{E}(Z) \leq \infty$. Let N be a stopping time, and *i.e.*, $\mathbb{E}(N) \leq \infty$. Then

$$\mathbb{E}(Z_1 + \cdots + Z_N) = \mathbb{E}(N)\mathbb{E}(Z)$$

Proof: Rewrite $Z_1 + \cdots + Z_N = \sum_{i=1}^{\infty} Z_i 1_{N \geq i}$. By the dominated convergence theorem

$$\mathbb{E}(Z_1 + \cdots + Z_N) = \sum_{i=1}^{\infty} \mathbb{E}(Z_i 1_{N \geq i})$$

Since N is a stopping time, $N \geq i$ defined by $\{Z_1, \cdots, Z_{i-1}\}$, the event $N \geq i$ is independent of Z_i. Therefore

$$\mathbb{E}(Z_1 + \cdots + Z_N) = \sum_{i=1}^{\infty} \mathbb{E}(Z_i)\mathbb{E}(1_{N \geq i}) = \mathbb{E}(Z) \sum_{i=1}^{\infty} \Pr[N \geq i]$$

Average Sample Size

We use the approximation that, when the terminal detection is made, with $Z_i = \log L(Y_i)$,

$$Z_1 + \cdots + Z_N \approx \begin{cases} \log B & \text{Choose } \mathcal{H}_1 \\ \log A & \text{Choose } \mathcal{H}_0 \end{cases}$$

By the Wald's equation, since Z_i are i.i.d.,

$$\mathbb{E}_{\theta_0}(N) \approx \frac{\alpha \log B + (1-\alpha) \log A}{\mathbb{E}_{\theta_0}(Z)}$$

$$\mathbb{E}_{\theta_1}(N) \approx \frac{\beta \log B + (1-\beta) \log A}{\mathbb{E}_{\theta_1}(Z)}$$

Example

Consider $\mathcal{H}_i : Y_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(i,1)$. Compare the average sample size the fixed sample size likelihood ratio detector and that of the SPRT under $\alpha = 1-\beta = Q^{-1}(4) = 3.16 \times 10^{-5}$.

SPRT For SPRT(A,B), we choose

$$\ln A = \ln \frac{1-\beta}{1-\alpha} = -\ln 9. \qquad \ln B = \ln \frac{\beta}{\alpha} = \ln 9.$$

The log-likelihood ratio for a single sample Y is

$$\begin{aligned} Z &= \ln L(Y) = \ln \frac{\exp\{-\frac{(Y-1)^2}{2}\}}{\exp\{-\frac{Y^2}{2}\}} \\ &= Y - \frac{1}{2} \sim \begin{cases} \mathcal{N}(-\frac{1}{2},1) & \text{under } \mathcal{H}_0 \\ \mathcal{N}(\frac{1}{2},1) & \text{under } \mathcal{H}_1 \end{cases} \end{aligned}$$

The average sample size for the two hypotheses are

$$\mathbb{E}_{\theta_0}(N) \approx \frac{\alpha \ln \frac{\beta}{\alpha} + (1-\alpha)\ln \frac{1-\beta}{1-\alpha}}{\mathbb{E}_{\theta_0}(Z)} \approx 20.71$$

Similarly, we have $\mathbb{E}_{\theta_1}(N) \approx 20.71$.

Fixed Size Likelihood Ratio Detector

The the minimum sample size for the fixed sample size detector must satisfy

$$n_{\text{FSS}} \geq [Q^{-1}(\alpha) - Q^{-1}(\beta)]^2 = 4[Q^{-1}(\alpha)]^2 \approx 64.$$

An Auxiliary Bayesian Problem

Hypotheses with Infinite Samples

Consider simple binary hypotheses

$$\mathcal{H}_0 \ : \ Y_k \overset{\text{i.i.d.}}{\sim} f(y|\theta_0)$$
$$\mathcal{H}_1 \ : \ Y_k \overset{\text{i.i.d.}}{\sim} f(y|\theta_1)$$

Costs and Risks

Let w_0 be the cost of false alarm, w_1 the cost for miss detection, and the cost for taking one sample is c. For a given prior $\pi = \Pr(\Theta = \theta_0]$, the Bayesian risk of detector δ with size α and power β is given by

$$r(\pi, \delta) = \pi(\alpha w_0 + cE_{\theta_0}(N)) + (1-\pi)((1-\beta)w_1 + cE_{\theta_1}(N))$$

The Bayesian Sequential Detector

The Bayesian detector δ_B is given by

$$\inf_{\delta} r(\pi, \delta).$$

where δ is chosen among sequential detectors with stopping and terminal decision rules.

Concavity of Bayesian Risks

Lemma Let $V(\pi)$ be the minimum risk among all detectors taking at least one sample, *i.e.*,

$$V(\pi) = \inf_{\delta \in \mathcal{D}} r(\pi, \delta)$$

where $\mathcal{D}$ is the set of detectors taking at least one sample. Then $V(\pi)$ is concave and $V(0) = V(1) = c$

Proof: Consider priors π, π' and $\pi'' = \lambda\pi + (1-\lambda)\pi'$ for any $\lambda \in (0,1)$.

$$r(\pi", \delta) = \lambda r(\pi, \delta) + \lambda' r(\pi', \delta)$$

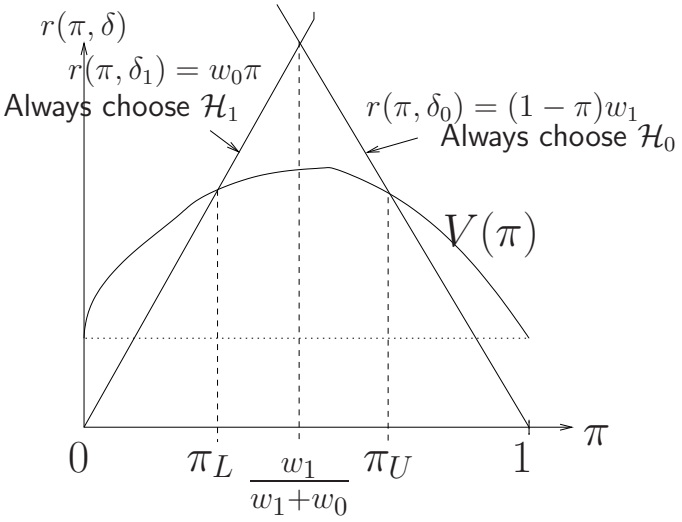
Thus

$$V(\pi'') = \inf_{\delta \in \mathcal{D}} r(\pi'', \delta) \geq \lambda \inf_{\delta \in \mathcal{D}} r(\pi, \delta) + \lambda' \inf_{\delta \in \mathcal{D}} r(\pi', \delta) = \lambda V(\pi) + \lambda' V(\pi')$$

For $\pi = 0$, the optimal sequential detector is to collect no more sample and set $\delta(y) = 1$, which gives $N = 1$ and $V(0) = c$.

Stopping Criterion

The straight lines are the risk when no sample is taken, and the curvy function is the risk when at least one sample is taken. Therefore, samples should be taken only if $\pi \in (\pi_L, \pi_U)$.



The Bayesian Detector

Lemma Given costs w_i, c, let $V(\pi) = r(\pi, \delta_B)$. Let $\pi_L, \pi_U \in [0,1]$ be such that

$$\begin{cases} V(\pi_L) = w_0\pi_L, \quad V(\pi_U) = w_1(1-\pi_U) & V(\frac{w_1}{w_1+w_0}) < \frac{w_1 w_0}{w_1+w_0} \\ \pi_L = \pi_U = \frac{w_1}{w_1+w_0} & o.w. \end{cases}$$

If $0 < \pi_L < \pi_U < 1$, then for any $\pi \in [\pi_L, \pi_U]$, the Bayesian detector that minimizes $r(\pi, \delta)$ is given by SPRT(A,B) with boundaries

$$A = \frac{\pi}{1-\pi}\frac{1-\pi_U}{\pi_U}, \quad B = \frac{\pi}{1-\pi}\frac{1-\pi_L}{\pi_L}$$

Conversely, for any $0 < \pi_L < \pi_U < 1$, there exist costs $w \in (0,1)$ and $c > 0$ such that the Bayes detector δ_B for the costs $w_0 = 1-w$, $w_1 = w$, and a prior $\pi \in (\pi_L, \pi_U)$ is an SPRT(A,B) with the same boundaries above.

Proof: (1) The direct part. Suppose that $\pi \in (\pi_L, \pi_U)$, and we have taken n observations $\mathbf{Y}_n = \mathbf{y}_n$. The probability that $\mathcal{H}_0$ is true is no longer π but the a posteriori probability

$$\pi(\mathbf{y}_n) = \frac{\pi f(\mathbf{y}_n|\theta_0)}{\pi f(\mathbf{y}_n|\theta_0) + (1-\pi) f(\mathbf{y}_n|\theta_1)}$$

Therefore, terminal decision should be made unless $\pi(\mathbf{y}_n) \in (\pi_L, \pi_U)$, which gives $L(\mathbf{y}_n) \in (A, B)$.

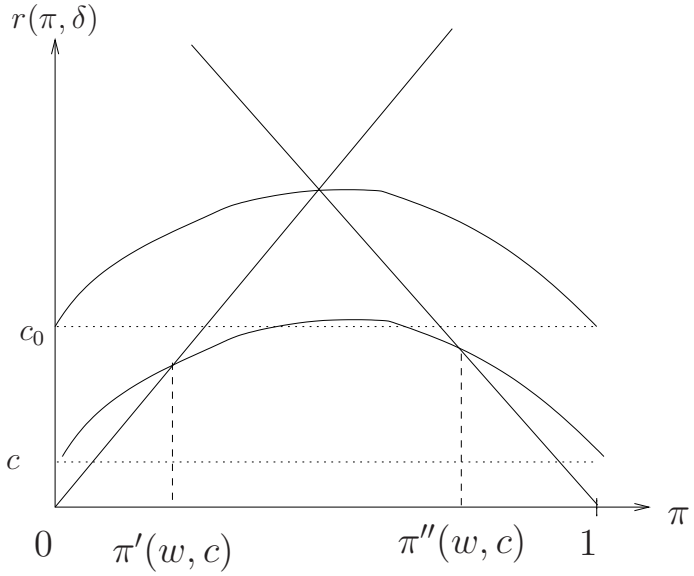
(2) Converse. To prove the converse, we note from the direct part that π_L and π_U are functions of w_i and c. The proof of converse consists of solving for costs w and c from given π_L and π_U. The proof includes the following steps

Bayesian risk as a function of c

Fix w, the Bayesian risk $V(\pi, w, c)$ is a continuous function of c.

- There exists c_0 such that $V(\pi, w, c) = (1-w)w$.
- $\pi'(w,c)$ and $\pi''(w,c)$ are continuous functions of c satisfying

$$\begin{cases} \pi'(w,c) \overset{c\to 0}{\to} 0 \\ \pi''(w,c) \overset{c\to 0}{\to} 1 \end{cases} \quad \pi''(w,c) - \pi'(w,c) \overset{c\to c_0}{\to} 0$$



Solving for costs

Given π_L and π_U, we want to solve for w and c from

$$\pi_L = \pi'(w,c), \quad \pi_U = \pi''(w,c)$$

Consider the following 1-1 transforms

$$\begin{aligned} \lambda_0 &= \frac{\pi_L}{1-\pi_L}\frac{1-\pi_U}{\pi_U}, \gamma_0 = \frac{\pi_U}{1-\pi_U} \\ \lambda(w,c) &= \frac{\pi'(w,c)}{1-\pi'(w,c)}\frac{1-\pi''(w,c)}{\pi''(w,c)}, \gamma(w,c) = \frac{\pi''(w,c)}{1-\pi''(w,c)}. \end{aligned}$$

Solving w and c is equivalent to solving w and c from

$$\gamma_0 = \gamma(w,c), \lambda_0 = \lambda(w,c)$$

Continuity Arguments

For fixed w, since $\pi'(w,c)$ and $\pi''(w,c)$ are continuous with respect to c, there exists $c(w)$ such that

$$\lambda_0 = \frac{\pi'(w,c(w))}{1-\pi'(w,c(w))}\frac{1-\pi''(w,c(w))}{\pi''(w,c(w))}.$$

We now only need to solve w from $\gamma(w,c(w)) = \gamma_0$.

Solving for w

First, consider costs w, $c(w)$ and prior $\pi = \pi'(w,c(w))$. By the direct part, there exists a Bayesian detector δ' that is SPRT(A',B') where

$$A' = \frac{\pi'(w,c(w))}{1-\pi'(w,c(w))}\frac{1-\pi''(w,c(w))}{\pi''(w,c(w))} = \gamma_0, \quad B' = 1$$

This Bayesian detector has false alarm α', power β', and average sample sizes $\mathbb{E}'_{\theta_i}(N)$, all are functions of γ_0 only. The Bayesian risk of δ' is given by

$$\begin{aligned}\pi'(w,c)(1-w) &= \pi'(w,c)\alpha'(1-w) + (1-\beta')(1-\pi'(w,c))w \\ &\quad + c[\mathbb{E}'_{\theta_0}(N)\pi'(w,c) + (1-\pi'(w,c))\mathbb{E}'_{\theta_1}(N)] \qquad (1)\end{aligned}$$

Second, consider costs w, $c(w)$ and prior $\pi = \pi''(w,c(w))$. By the direct part, there exists a Bayesian detector δ'' that is SPRT(A'',B'') where

$$A'' = 1, \quad B'' = \frac{1}{\gamma_0}$$

Again, this Bayesian detector has false alarm α'', power β'', and average sample sizes $\mathbb{E}''_{\theta_i}(N)$, and all are functions of γ_0 only. The Bayesian risk of δ'' is given by

$$\begin{aligned}(1-\pi''(w,c))w &= \pi''(w,c)\alpha''(1-w) + (1-\beta'')(1-\pi''(w,c))w \\ &\quad + c[\mathbb{E}''_{\theta_0}(N)\pi''(w,c) + (1-\pi''(w,c))\mathbb{E}''_{\theta_1}(N)] \qquad (2)\end{aligned}$$

Third, by definition

$$\frac{\pi'(w,c)}{1-\pi'(w,c)} = \lambda_0\gamma_0, \quad \frac{\pi''(w,c)}{1-\pi''(w,c)} = \gamma_0.$$

Substituting the above into (1)-(2), we have

$$\begin{aligned}\lambda_0\gamma_0(1-w) &= \lambda_0\gamma_0\alpha'(1-w) + (1-\beta')w + c[\lambda_0\gamma_0\mathbb{E}'_{\theta_0}(N) + \mathbb{E}'_{\theta_1}(N)] \\ w &= \gamma_0\alpha''(1-w) + (1-\beta'')w + c[\gamma_0\mathbb{E}''_{\theta_0}(N) + \mathbb{E}''_{\theta_1}(N)]\end{aligned}$$

From the above two equations, we solve for w and $c(w)$ given γ_0 and λ_0.

Proof of The Wald-Wolfowitz Theorem

The SPRT Detector Consider any SPRT(A,B) detector with $A < 1 < B$. For any constant $\pi \in (0,1)$, let

$$0 < \pi_L \triangleq \frac{\pi}{B(1-\pi)+\pi} < \pi_U \triangleq \frac{\pi}{A(1-\pi)+\pi} < 1$$

By the Lemma, SPRT(A,B) is the Bayesian detector for some $w_0 = 1-w$, $w_1 = w$, and per sample cost c. Let α_s be the size of SPRT(A,B) and β_s the power, and $\mathbb{E}_{\theta_i}(N_s)$ the average sample size.

Alternative Detector Consider any other detector δ with size $\alpha \leq \alpha_s$, $\beta \geq \beta_s$, and sample sizes $\mathbb{E}_{\theta_i}(N) < \infty$.

Comparing Performance Since SPRT(A,B) minimizes the Bayesian risk

$$\begin{aligned}&\pi[w_0\alpha_s + c\mathbb{E}_{\theta_0}(N_s)] + (1-\pi)[w_1(1-\beta_s) + c\mathbb{E}_{\theta_1}(N_s)] \\ &\leq \pi[w_0\alpha + c\mathbb{E}_{\theta_0}(N)] + (1-\pi)[w_1(1-\beta) + c\mathbb{E}_{\theta_1}(N)] \qquad (3)\end{aligned}$$

Because $\alpha \leq \alpha_s$ and $\beta \geq \beta_s$, we have

$$\pi\mathbb{E}_{\theta_0}(N_s) + (1-\pi)\mathbb{E}_{\theta_1}(N_s) \leq \pi\mathbb{E}_{\theta_0}(N) + (1-\pi)\mathbb{E}_{\theta_1}(N)$$

Since the above is true for any π, we have

$$\mathbb{E}_{\theta_i}(N_s) \leq \mathbb{E}_{\theta_i}(N)$$