The characteristic function of the gamma $(\frac{1}{2}, \frac{1}{2\sigma_{jk}^2})$ density is

$$\phi_{T_k}(u) = \frac{1}{[1-2iu\sigma_{jk}^2]^{\frac{1}{2}}}, \quad u \in \mathbb{R}$$

$$\Rightarrow \quad p_T(t|H_j) = \int_{-\infty}^{\infty} e^{-iut} \prod_{k=1}^{n} [1-2iu\sigma_{jk}^2]^{-\frac{1}{2}} du$$

When $\sigma_{j1}^2 = \cdots = \sigma_{jn}^2 = \sigma_j^2$,

$$p_T(t|H_j) = \begin{cases} \frac{1}{(2\sigma_j^2)^{n/2}\Gamma(\frac{n}{2})} t^{\frac{n}{2}-1} e^{-\frac{t}{2\sigma_j^2}}, & t> \\ 0, & t\le \end{cases}$$

where $\Gamma(x) = \int_0^{\infty} e^{-y} y^{x-1} dy$ is the gamma function.

This case corresponds to the situation when $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \sigma_s^2$ and

$$\Sigma_s = \sigma_s^2 I.$$

$$\Rightarrow \quad P_j(\underline{Y}^T Q \underline{Y} > \tau') = 1 - \Gamma(\frac{n}{2}; \frac{\tau'}{2\sigma_j^2})$$

where $\Gamma(x;t) = \int_0^t e^{-y} y^{x-1} dy / \Gamma(x)$ is the incomplete gamma function.

$$P_j(\underline{Y}^T Q \underline{Y} > \tau') = e^{-\tau'/(2\sigma_j^2)} \sum_{k=0}^{\frac{n}{2}-1} \frac{(\tau'/(2\sigma_j^2))^k}{k!}$$

* For Neyman-Pearson detection with false-alarm probability α, we choose

$$\tau' = 2\sigma_0^2 \Gamma^{-1}\left(\frac{n}{2}; 1-\alpha\right)$$

where $\Gamma^{-1}(x; \cdot)$ is the inverse function of $\Gamma(x; \cdot)$ in its second variable. The ROCs are given by

$$P_D(\tilde{\delta}_{NP}) = 1 - \Gamma\left[\frac{n}{2}; \frac{\sigma_0^2}{\sigma_1^2} \Gamma^{-1}\left(\frac{n}{2}; 1-\alpha\right)\right]$$

$\Rightarrow$ The performance is parameterized by two parameters n and

$$\frac{\sigma_0^2}{\sigma_1^2} = \frac{1}{1 + \sigma_s^2/\sigma^2}, \quad \frac{\sigma_s^2}{\sigma^2} \text{ is the SNR}$$

When the signal is not i.i.d., the solution does not have a closed-form.

* <u>Remark 1</u>: A Relationship Between the Dependent and Independent Signals Cases

When $\underline{N} \sim N(\underline{0}, \sigma^2 I)$, $\underline{S} \sim N(\underline{\mu}, \Sigma_s)$ and Σ_s is diagonal,
$\Sigma_s = \operatorname{diag}(\sigma_{s_1}^2, \ldots, \sigma_{s_n}^2)$

$$\log L(\underline{y}) = \frac{1}{2}\sum_{k=1}^{n} y_k^2/\sigma^2 - \frac{1}{2}\sum_{k=0}^{n} (y_k - \mu_k)^2/(\sigma_{s_k}^2 + \sigma^2)$$
$$+ \frac{1}{2}\sum_{k=1}^{n} \log\left[\sigma^2/(\sigma_{s_k}^2 + \sigma^2)\right]$$

Next, consider Σ_s not diagonal. $\underline{Y}$ is a multivariate Gaussian. Let $p_j(y_1, \ldots, y_\ell)$ denote the joint density of $Y_1, \ldots, Y_\ell$ under H_j. Then,

$$p_j(\underline{y}) = p_j(y_1) \prod_{k=2}^{n} p_j(y_k | y_1, \ldots, y_{k-1})$$

Where $p_j(y_k / y_1, \ldots, y_{k-1})$ is the conditional density of y_k given $Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}$.

Under H_0, Y_k is $Y_1, \ldots, Y_{k-1}$ since $\underline{N}$ is i.i.d. So,

$$p_0(\underline{y}) = \Pi_{k=1}^{n} \; p_0(y_k).$$

Under H_1, Y_k is not independent of $Y_1, \ldots, Y_{k-1}$ and Y_k is conditional Gaussian given $Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}$. The mean of this conditional density is given by

$$E_1\{Y_k / Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}\}$$

$$= E_1\{S_k / Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}\} + E_1\{N_k / Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}\}$$
$$= E_1\{S_k / Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}\} \triangleq \hat{S}_k$$

(N_k is independent of $Y_1, \ldots, Y_{k-1}$ and has 0 mean)

The variance of the conditional density is

$$Var_1(Y_k | Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1})$$
$$= Var_1(S_k | Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1})$$
$$+ Var_1(N_k | Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1}) = \hat{\sigma}_{S_k}^2 + \sigma^2$$

where $\hat{\sigma}_{S_k}^2 \stackrel{\Delta}{=} Var_1(S_k \mid Y_1 = y_1, \ldots, Y_{k-1} = y_{k-1})$

$$\Rightarrow \quad \hat{s}_1 = E(S_1), \quad \hat{\sigma}_{S_1}^2 = Var(S_1)$$

$$p_1(\underline{y}) = \prod_{k=1}^{n} \mathcal{N}(\hat{s}_k, \hat{\sigma}_{S_k}^2 + \sigma^2)$$

$$\Rightarrow \log(\underline{y}) = \frac{1}{2} \sum_{k=1}^{n} y_k^2/\sigma^2 - \frac{1}{2} \sum_{k=1}^{n} (y_k - \hat{s}_k)^2/(\hat{\sigma}_{S_k}^2 + \sigma^2)$$
$$+ \frac{1}{2} \sum_{k=1}^{n} \log\left[\sigma^2/(\hat{\sigma}_{S_k}^2 + \sigma^2)\right]$$

$\Rightarrow$ Detecting a dependent stochastic signal is analogous to detecting an independent stochastic signal with mean $\hat{\underline{s}}$ and covariance $\mathrm{diag}(\hat{\sigma}_{S_1}^2, \ldots, \hat{\sigma}_{S_n}^2)$

The difference is that $\hat{s}_k$ depends on $y_1, \ldots, y_{k-1}$.

Another way to view the above:

under H_1,

$$Y_k = N_k + S_k = N_k + \varepsilon_k + \hat{S}_k$$

where $\hat{S}_k$ is interpreted as the random quantity $E_1\{S_k \mid Y_1, \ldots, Y_{k-1}\}$ and $\varepsilon_k = S_k - \hat{S}_k$. As we will see later, $\hat{S}_k$ is an optimum predictor (under H_1) of S_k from the past observations $Y_1, \ldots, Y_{k-1}$. ε_k can be interpreted as the prediction error, or as the part of S_k that can not be predicted from the past observations.

⇒ For each observation, we can think of the signal as consisting of a part, $\hat{S}_k$, known from the past, and a new part ε_k that can not be predicted from the past.

⇒ It can be known that under H_1, ε_k is statistically independent of $Y_1, \cdots, Y_{k-1}$ and is a $N(0, \hat{\sigma}^2_{s_k})$ random variable.

When $S_1, \cdots, S_n$ are independent, $\hat{S}_k = \mu_k$ and $\varepsilon_k = S_k - \mu_k$ that is $N(0, \sigma^2_{S_k})$.

* <u>Remark 2</u>: Estimator — Correlator Interpretation of the Optimum Detector for Stochastic Signals in i.i.d. Gaussian Noise

Coming back to

$$\log L(\underline{y}) = \frac{1}{2\sigma^2}\left[\sum_{k=1}^{n} y_k^2 - \sum_{k=1}^{n} (y_k - \hat{S}_k)^2 / (1 + \hat{\sigma}^2_{S_k}/\sigma^2)\right] - \frac{1}{2}\sum_{k=1}^{n} \log\left(1 + \frac{\hat{\sigma}^2_{S_k}}{\sigma^2}\right)$$

Suppose the prediction error variance is small,

$$\hat{\sigma}^2_{S_k} \ll \sigma^2 \quad \text{for all } k.$$

Then $1 + \dfrac{\hat{\sigma}^2_{S_k}}{\sigma^2} \approx 1$ and

$$\log L(\underline{y}) \cong \frac{1}{\sigma^2}\left[\sum_{k=1}^{n} y_k \hat{S}_k - \frac{1}{2}\sum_{k=1}^{n} (\hat{S}_k)^2\right]$$

This is the likelihood ratio for detecting $\hat{s}_1, \dots, \hat{s}_n$ as if it were coherent signal.

$\Rightarrow$ We can view the stochastic signal detector approximately as one that estimates the signal and then treat it as a known signal

In general, let a stochastic signal be with multivariate pdf $p_{\underline{s}}$ embedded in $\mathcal{N}(\underline{0}, \sigma^2 I)$ noise. The likelihood ratio is

$$L(\underline{y}) = \int_{\mathbb{R}^n} \exp\left\{\frac{1}{\sigma^2}\left(\underline{s}^T \underline{y} - \frac{1}{2}\|\underline{s}\|^2\right)\right\} p_{\underline{s}}(\underline{s})\, d\underline{s}$$

Assume the regularity of $p_{\underline{s}}$. By the mean-value theorem, we have

$$L(\underline{y}) = \exp\left\{\frac{1}{\sigma^2}\left(\hat{\underline{s}}^T \underline{y} - \frac{1}{2}\|\hat{\underline{s}}\|^2\right)\right\}$$

for some $\hat{\underline{s}} \in \mathbb{R}^n$ (here, of course, $\hat{\underline{s}}$ depends on $\underline{y}$).

$\Rightarrow$ The likelihood ratio for stochastic signals in i.i.d. Gaussian noise can be interpreted as an "estimator", $\hat{\underline{s}}$, of the signal followed by the optimum detector for $\hat{\underline{s}}$ as if it were a coherent signal. This structure is known as an <u>estimator correlator</u>.

However, the determination of $\hat{\underline{s}}(\underline{y})$ is difficult (studied later).

* Remark 3 : Locally Optimum Detection of Stochastic Signals

Consider the following composite hypothesis testing problem:

$$H_0 : Y_k = N_k, \quad k = 1, 2, \cdots, n$$

vs

$$H_1 : Y_k = N_k + \theta^{\frac{1}{2}} S_k, \quad k = 1, 2, \cdots, n, \; \theta > 0,$$

where $\underline{N}$ and $\underline{S}$ are the same as before.

The covariance of $\theta^{\frac{1}{2}} \underline{S}$ is $\theta \Sigma_s$ and the quadratic detection statistics for Neyman – Pearson testing with fixed θ is

$$\theta \, \underline{y}^T \Sigma_s (I + \theta \Sigma_s)^{-1} \underline{y}$$

This θ can be absorbed into the decision threshold

can not be decoupled from the observations $\underline{y}$

$\Downarrow$

No UMP test exists.

An LMP test statistic can be found by taking the derivative of the above statistics in terms of θ:

$$2 \, \underline{y}^T \Sigma_s \underline{y}$$

A special case for Σ_s when the $k-l$th element of Σ_s, say $\rho_{k,l}$, depends only on the difference

$(k-l)$. In this case, $\rho_{k,l} = \rho_{k-l,0} \triangleq \rho_{k-l}$.

A signal with this property is said to be wide-sense stationary.

Consider a scaled but equivalent LMP statistic:

$$T(\underline{y}) = \frac{1}{n} \underline{y}^T \Sigma_S \underline{y}$$

$$= \frac{1}{n} \sum_{k=1}^{n} \sum_{l=1}^{n} y_k y_l \rho_{k-l}$$

$$= \rho_0 \hat{\rho}_0 + 2 \sum_{k=1}^{n-1} \rho_k \hat{\rho}_k$$

where $\hat{\rho}_k \triangleq \frac{1}{n-k} \sum_{l=1}^{n-k} y_l y_{l+k}$, $k = 0, 1, \cdots, n-1$.

$\Rightarrow$ When $n \gg k$, $\hat{\rho}_k$ is an estimate of the covariance $E\{Y_l Y_{l+k}\}$ for $l = 1, 2, \cdots, n-k$.

$\Rightarrow$ $T(\underline{y})$ can be interpreted as: first estimate the covariance sequence from the observations and then correlates it with the signal covariance sequence.

Under H_0,

$$E\{Y_l Y_{l+k}\} = \begin{cases} 1, & \text{if } k=0 \\ 0, & \text{if } k \neq 0 \end{cases}$$

Under H_1,

$$E\{Y_l Y_{l+k}\} = \begin{cases} 1+\theta\rho_0, & \text{if } k=0 \\ \theta \rho_k, & \text{if } k \neq 0. \end{cases}$$

Assume that the estimates $\hat{\rho}_k$ were reasonablly accurate, we would have

$$T(\underline{y}) \cong \begin{cases} \rho_0 & \text{under } H_0 \\ \rho_0 + \theta(\rho_0^2 + 2\sum_{k=1}^{n-1} \rho_k^2) & \text{under } H_1 \end{cases}$$

↳ looks reasonable to detect a signal.

* Frequency domain interpretation for $T(\underline{y})$

Let us think that $S_1, \cdots, S_n$ as a segment of an infinite random sequence $\{S_k\}_{k=-\infty}^{\infty}$ with $E\{S_l S_{l+k}\} = \rho_k$ for all l and k.

Let $\phi(\omega) \triangleq \sum_{k=-\infty}^{\infty} \rho_k e^{-i\omega k}$

that is the power spectrum of $\{S_k\}_{k=-\infty}^{\infty}$.

$\Rightarrow T(\underline{y}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(\omega) \hat{\phi}(\omega) \, d\omega$

where $\hat{\phi}(\omega) = \frac{1}{n} \left| \sum_{k=1}^{n} y_k e^{i\omega k} \right|^2, \quad -\pi \le \omega \le \pi.$

is known as the periodogram of the data as an estimate of the spectrum of the observations.

$\Rightarrow$ $T(\underline{y})$ first estimates the observation spectrum then correlates this estimate with the signal spectrum.

§2. Performance Evaluation of Signal Detection Procedures

In general it is hard to have closed-form solution for the false-alarm and miss probabilities:

$$P_F(\tilde{\delta}) = P_0(\tilde{\delta} \text{ chooses } H_1)$$

$$P_M(\tilde{\delta}) = P_1(\tilde{\delta} \text{ chooses } H_0)$$

for a decision rule $\tilde{\delta}$.

Likelihood-ratio tests and most other decision rules of interest are of the form

$$\tilde{\delta}_T(y) = \begin{cases} 1 & \text{if } T(y) > \tau \\ \gamma & \phantom{\text{if } T(y)} = \\ 0 & \phantom{\text{if } T(y)} < \end{cases}$$

where T is a mapping from $(\Gamma, \mathcal{G})$ to $(\mathbb{R}, \mathcal{B})$. The performance analysis is to compute the probabilities of the regions

$\{T(Y) > \tau\}$ (or $\{T(Y) < \tau\}$) and $\{T(Y) = \tau\}$

under the two hypotheses.

* Direct Performance Computation

$$P_F(\tilde{\delta}_T) = P(T(Y) > \tau \mid H_0) + \gamma (P(T(Y) = \tau \mid H_0)$$
$$= [1 - F_{T,0}(\tau)] + \gamma [F_{T,0}(\tau) - \lim_{\sigma \to \tau^-} F_{T,0}(\sigma)]$$

and

$$P_M(\tilde{\delta}_T) = P(T(Y)<\tau \mid H_1) + (1-\gamma)P(T(Y)=\tau \mid H_1)$$
$$= P(T(Y)\le \tau \mid H_1) - \gamma P(T(Y)=\tau \mid H_1)$$
$$= F_{T,1}(\tau) - \gamma\left[F_{T,1}(\tau) - \lim_{\sigma\to\tau^-} F_{T,1}(\sigma)\right]$$

where $F_{T,j}$ is the cumulative distribution function (cdf) of $T(Y)$ under hypothesis H_j.

* In one case $F_{T,j}$ can be calculated : when $\underline{Y} = (Y_1, \ldots, Y_n)^T$ is a vector of independent real random variables and $T(\underline{y})$ has the structure

$$T(\underline{y}) = \sum_{k=1}^{n} g_k(y_k)$$

where $\{g_k\}_{k=1}^{n}$ is a sequence of nonlinearities (for example, the log-likelihood ratio).

In this case, $F_{T,j}$ can be calculated by using characteristic functions. Let $\phi_{T,j}$ and $\phi_{g_k,j}$ be the characteristic functions of $T(\underline{Y})$ and $g_k(Y_k)$, respectively, under H_j.

Then,
$$\phi_{T,j}(u) = E\left\{\exp\left\{iu\sum_{k=1}^{n} g_k(Y_k)\right\} \middle| H_j\right\}$$
$$= \prod_{k=1}^{n} E\{\exp\{iu\, g_k(Y_k)\} \mid H_j\}$$
$$= \prod_{k=1}^{n} \phi_{g_k,j}(u)$$

$$\Rightarrow F_{T,j}(b) - F_{T,j}(a) = \lim_{U\to\infty} \frac{1}{2\pi}\int_{-U}^{U} \frac{e^{-iua} - e^{-iub}}{iu}\, \phi_{T,j}(u)\, du$$

for all a and b that are continuity points of $F_{T,j}$

When $T(\underline{Y})$ is a continuous random variable under H_j,

$$p_{T,j}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_{T,j}(u)\, e^{-iut}\, du$$

$$\Rightarrow P_F(\tilde{\delta}_T) = \frac{1}{2\pi}\int_{\tau}^{\infty}\int_{-\infty}^{\infty} e^{-iut}\left[\prod_{k=1}^{n}\phi_{g_{k,0}}(u)\right] du\, dt$$

$P_M(\tilde{\delta}_T)$ is similar.

* Example: Correlation Detection in Cauchy Noise

Consider the performance of the correlation detector

$$g_k(y_k) = s_k y_k, \quad k = 1, 2, \cdots, n.$$

The hypothesis pair is

$$H_0 : Y_k = N_k, \quad k = 1, 2, \cdots, n$$

v.s.

$$H_1 : Y_k = N_k + S_k, \quad k = 1, 2, \cdots, n,$$

where $N_1, \cdots, N_n$ is i.i.d. with Cauchy pdf

$$p_{N_k}(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.$$

$$\Rightarrow \phi_{g_{k,0}}(u) = E\{e^{iu s_k N_k}\} = \phi_{N_k}(u s_k)$$

$$= \mathcal{F}\{p_{N_k}\}\Big|_{u s_k} = e^{-|s_k u|}, \quad u \in \mathbb{R}.$$

$$\Rightarrow \phi_{T,0}(u) = \prod_{k=1}^{n} \phi_{g_{k,0}}(u) = e^{-n|\bar{s}||u|}$$

where $|\bar{s}| = \frac{1}{n} \Sigma_{k=1}^{n} |s_k|$.

$$\Rightarrow P_F(\tilde{\delta}_T) = \frac{1}{2\pi} \int_{\tau}^{\infty} \int_{-\infty}^{\infty} e^{-iut} e^{-n|\bar{s}||u|} \, du\,dt$$

$$= \frac{1}{n|\bar{s}|\pi} \int_{\tau}^{\infty} \frac{1}{1+\left(\frac{t}{n|\bar{s}|}\right)^2} \, dt$$

$$= \frac{1}{2} - \frac{1}{\pi} \mathrm{Tan}^{-1}\left(\frac{\tau}{n|\bar{s}|}\right)$$

Similarly, $\phi_{T,1}(u) = \phi_{T,0}(u)\, e^{iun\overline{s^2}}$

where $\overline{s^2} = \frac{1}{n} \sum_{k=1}^{n} s_k^2$

$$\Rightarrow P_M(\tilde{\delta}_T) = \frac{1}{2} + \frac{1}{\pi} \mathrm{Tan}^{-1}\left(\frac{\tau}{n|\bar{s}|} - \sqrt{\overline{s^2}}\right).$$

We can achieve size-α detection by choosing

$$\tau = n|\bar{s}| \tan\left(\tfrac{1}{2} - \alpha\right)$$

and the ROCs is

$$P_D(\tilde{\delta}_T) = 1 - P_M(\tilde{\delta}_T)$$

$$= \frac{1}{2} - \frac{1}{\pi} \mathrm{Tan}^{-1}\left(\tan\left(\tfrac{1}{2} - \alpha\right) - \sqrt{\overline{s^2}}\right).$$

$\Rightarrow$ Surprisingly, the above detection probability tells us that the performance is improved by increasing the average signal power but <u>NOT</u>

by the number of samples.

This odd behavior is due to the fact that the correlation detector is quite different from the optimum detector for coherent signals in Cauchy noise, for which performance does improve with increased sample size. The heaviness of the tails of the Cauchy distribution cancels the effect of noise reduction that the correlator achieves in the Gaussian case.

* Chernoff and Related Bounds

In most cases, it is hard to have exact error probabilities P_F and P_M and then good bounds on them may be good enough.

Markov Inequality: Suppose that X is a random variable. If $P(X \geq 0) = 1$, then

$$P(X \geq a) \leq E\{X\}/a \quad \text{for all } a > 0.$$

Proof: $P(X \geq a) = E\{I_{[a,\infty)}(X)\}$ where $I_{[a,\infty)}$ is the indicator function of the set $[a,\infty)$:

$$I_{[a,\infty)}(x) = \begin{cases} 1 & \text{if } x \geq a, \\ 0 & \text{if } x < a. \end{cases}$$

Since $\bar{\gamma} \geq 0$,

$$I_{[a,\infty)}(\bar{\gamma}) \leq \frac{\bar{\gamma}}{a} \Rightarrow E\{I_{[a,\infty)}(\bar{\gamma})\} \leq \frac{E\{\bar{\gamma}\}}{a} \quad \square$$

Since in the randomization, $\gamma \leq 1$,

$$P_F(\tilde{\delta}_T) \leq P_0(T(Y) \geq \tau) = P_0(e^{sT(Y)} \geq e^{s\tau})$$

Markov inequality $\qquad \leq e^{-s\tau} E\{e^{sT(Y)} | H_0\}$

$$= \exp\{-s\tau + \mu_{T,0}(s)\}, \text{ for all } s \geq 0$$

where $\mu_{T,0}$ is the cumulant generating function (cgf) of $T(Y)$ under H_0:

$$\mu_{T,0}(s) = \log(E\{e^{sT(Y)} | H_0\})$$

Similarly, since $\gamma \geq 0$, we have

$$P_M(\tilde{\delta}_T) \leq P_1(T(Y) \leq \tau) = P_1(e^{tT(Y)} \geq e^{t\tau})$$

$$\leq \exp\{-t\tau + \mu_{T,1}(t)\}$$

for all $t \leq 0$, where $\mu_{T,1}$ is the cgf of $T(Y)$ under H_1.

The above two bounds can be minimized over $s > 0$ and $t < 0$ to the tighest such bounds provided the cgf's of $T(Y)$ are known.

For the likelihood ratio detector, let us assume that P_j has density p_j for $j=0$ and 1, and $T(y) = \log L(y)$, where $L(y) = p_1(y)/p_0(y)$. Then,

$$\mu_{T,0}(s) = \log\left(\int_\Gamma e^{s\log L}\, p_0\, d\mu\right)$$

$$= \log\left(\int_\Gamma L^s\, p_0\, d\mu\right)$$

and $\mu_{T,1}(t) = \log\left(\int_\Gamma L^t\, p_1\, d\mu\right)$

$$= \log\left(\int_\Gamma L^{t+1}\, p_0\, d\mu\right)$$

$$= \mu_{T,0}(t+1)$$

$\Rightarrow$ $P_M(\delta_T) \le \exp\{(1-s)\tau + \mu_{T,0}(s)\}$, for $s<1$,

Compare it to $P_F(\delta_T) \le \exp\{-s\tau + \mu_{T,0}(s)\}$, for $s>0$.

Both of the bounds achieve their minima at the same value of s if

$$\arg\left\{\min_{s<1} |\mu_{T,0}(s) - s\tau|\right\} > 0$$

and $\arg\left\{\min_{s>0} |\mu_{T,0}(s) - s\tau|\right\} < 1.$

It can be shown that $\mu_{T,0}(s) - s\tau$ is a convex function of s on the region where

$\mu_{T,0}(s) < \infty$.

$\Rightarrow \mu'_{T,0}(s) = \dfrac{d\mu_{T,0}(s)}{ds} = \tau$

is sufficient for a minimum of $\mu_{T,0}(s) - s\tau$.

Let $\mu'_{T,0}(s_0) = \tau$ for $0 < s_0 < 1$.

Then, Chernoff bounds

$$P_F(\tilde{\delta}_T) \le \exp\{\mu_{T,0}(s_0) - s_0\,\mu'_{T,0}(s_0)\}$$

$$P_M(\tilde{\delta}_T) \le \exp\{\mu_{T,0}(s_0) + (1-s_0)\,\mu'_{T,0}(s_0)\}$$

$\Rightarrow$ If priors π_0 and π_1 are known, the average probability of error is bounded:

$$P_e = \pi_0 P_F + \pi_1 P_M$$

$$\le [\pi_0 + \pi_1 e^{\mu'_{T,0}(s_0)}]\exp\{\mu_{T,0}(s_0) - s_0\mu'_{T,0}(s_0)\}$$

$$\Rightarrow P_e \le \pi_0 e^{-s\tau}\int_{\Gamma_1} L^s p_0\,d\mu + \pi_1 e^{(1-s)\tau}\int_{\Gamma_0} L^s p_0\,d\mu$$

for $0 \le s \le 1$, where $\Gamma_1 = \{L(y) \ge \tau\}$, $\Gamma_0 = \Gamma_1^c$.

$$\Rightarrow P_e \le \max\{\pi_0, \pi_1 e^{\tau}\}\exp\{\mu_{T,0}(s) - s\tau\}$$

$0 \le s \le 1$.

$\Rightarrow$ with $\tau = \log\frac{\pi_0}{\pi_1}$

$$P_e \le \pi_0^{1-s}\,\pi_1^{s}\,e^{\mu_{T,0}(s)}, \quad 0 \le s \le 1.$$

* **Example: The Chernoff Bound for Quadratic Detection**

To illustrate the use of the Chernoff bound by considering the problem of detecting a $\mathcal{N}(\underline{0}, \Sigma_s)$ signal in $\mathcal{N}(\underline{0}, \sigma^2 I)$ noise.

After transforming $Y_1, \cdots, Y_n$ into the independent sequence $\bar{Y}_1, \cdots, \bar{Y}_n$, we have

$$L(\underline{y}) = \prod_{k=1}^{n} \frac{\sigma_{0k}}{\sigma_{1k}} e^{\bar{y}_k^2}$$

with $\sigma_{jk}^2 = Var(\bar{Y}_k \mid H_j) = \begin{cases} \frac{\lambda_k}{\sigma^2 + \lambda_k}, & \text{if } j=0 \\ \frac{\lambda_k}{\sigma^2}, & \text{if } j=1 \end{cases}$

$$\Rightarrow \mu_{T,0}(s) = \log\left(E\left\{\prod_{k=1}^{n} \left(\frac{\sigma_{0k}}{\sigma_{1k}}\right)^s e^{s\bar{Y}_k^2} \,\middle|\, H_0\right\}\right)$$

$$= \sum_{k=1}^{n} s \log\frac{\sigma_{0k}}{\sigma_{1k}} + \sum_{k=1}^{n} \log\left(E\{e^{s\bar{Y}_k^2} \mid H_0\}\right)$$

The expectation there is

$$E\{e^{s\bar{Y}_k^2} \mid H_0\} = \begin{cases} (1 - s\,\sigma_{0k}^2)^{-\frac{1}{2}} & \text{if } s < \frac{1}{\sigma_{0k}^2} \\ \infty & \text{if } s \geq \frac{1}{\sigma_{0k}^2} \end{cases}$$

Since $\sigma_{0k}^2 < 1$ for all k, they are finite for all k when $s \leq 1$. Thus

$$P_e \leq \pi_0^{1-s} \pi_1^{s} \prod_{k=1}^{n} \frac{\sigma^s (\sigma^2 + \lambda_k)^{(1-s)/2}}{[\sigma^2 + (1-s)\lambda_k]^{1/2}}, \quad 0 \leq s \leq 1.$$

The bound is minimized by the value s_0 solving

$$2\log\frac{\pi_0}{\pi_1}+\sum_{k=1}^{n}\log\left(1+\frac{\lambda_k}{\sigma^2}\right)=\sum_{k=1}^{n}\frac{\lambda_k}{\sigma^2+(1-s_0)\lambda_k}$$

* Several other Bounds

Bhattacharyya bound: Take $s=\frac{1}{2}$

$$P_e \le \sqrt{\pi_0\pi_1}\exp\{\mu_{T,0}(\tfrac{1}{2})\}$$

where Bhattacharyya coefficient or the Hellinger integral (also the affinity)

$$\rho \triangleq \exp\{\mu_{T,0}(\tfrac{1}{2})\}=\int_\Gamma [p_0 p_1]^{1/2}\,d\mu.$$

Lower bounds [Kobayashi and Thomas '67]

$$\pi_0\pi_1\rho^2 \le P_e \le (\pi_0\pi_1)^{1/2}\rho$$

$$P_e \ge \pi_0\pi_1 e^{-J/2}$$

where $J = \int_\Gamma (L-1)\log(L)\, p_0\, d\mu$.
(J-divergence) (or relative entropy between p_1 and p_1).

The reason to study these bounds is because $E(L^s)$ and $E((L-1)\log L)$ are usually easier to compute than the error probability.

* Asymptotic Relative Efficiency

When exact error probabilities are hard to calculate, considering other criteria may be helpful too. The asymptotic relative effeciency (ARE) is a criterion for large sample size detections.

Observations $Y_1, Y_2, \cdots$ obey one of two statistical hypotheses, H_0 and H_1.
$\tilde{\delta}_1$ and $\tilde{\delta}_2$ are two tests for H_0 and H_1 that have identical error probabilities but use n_1 and n_2 samples, respectively. If $n_1 < n_2$, we might say that $\tilde{\delta}_1$ is more efficient than $\tilde{\delta}_2$ because it requires less information than $\tilde{\delta}_2$ does to achieve identical performance.

$\Rightarrow$ The ratio $\frac{n_2}{n_1}$ is a good measure of the efficiency of $\tilde{\delta}_1$ relative to $\tilde{\delta}_2$. This becomes particularly useful for large sample sizes (large n_1 and n_2), although both systems would probably achieve good error-probability performance with large numbers of samples.

Pitman's ARE is an asymptotic ($n_1, n_2 \to \infty$) measure of the efficiency of one detector relative to another as follows.

Consider two sequences $\{\tilde{\delta}_1^{(n)}\}_{n=1}^{\infty}$ and $\{\tilde{\delta}_2^{(n)}\}_{n=1}^{\infty}$ of tests of H_0 vs. H_1, where $\tilde{\delta}_j^{(n)}$ operates with n samples.

Assume that the false-alarm probability of each test in each sequence is fixed at $\alpha \in (0,1)$.

The relative efficiency of $\{\tilde{\delta}_1^{(n)}\}_{n=1}^{\infty}$ relative to $\{\tilde{\delta}_2^{(n)}\}$ for sample size n is defined as $\dfrac{n_2}{n}$

where n_2 is the smallest number of samples such that $P_D(\tilde{\delta}_2^{(n_2)}) \geq P_D(\tilde{\delta}_1^{(n)}) \triangleq \beta_n$.

We would like to define the ARE as the limit of the relative efficiency as n approaches ∞. However, for most reasonable test sequences, $\lim_{n\to\infty} \beta_n = 1$ (called "consistent"), for very large n, β_n is no longer suitable to use in the definition of the relative efficiency.

To overcome this difficulty, we consider a sequence of alternative hypotheses $\{H_1^{(n)}\}_{n=1}^{\infty}$ converging in someway to H_0 such that $\lim_{n\to\infty} \beta_n \in (\alpha, 1)$. We then compute the relative efficiency assuming that both $\tilde{\delta}_1^{(n)}$ and $\tilde{\delta}_2^{(n)}$ are tests of H_0 vs. $H_1^{(n)}$ and define the asymptotic efficiency of $\{\tilde{\delta}_1^{(n)}\}_{n=1}^{\infty}$ relative to $\{\tilde{\delta}_2^{(n)}\}_{n=1}^{\infty}$ by

$$ARE_{1,2} = \lim_{n\to\infty} \frac{n_2}{n}$$

Note that $H_1^{(n)}$ being "close" to H_0 (for large n) corresponds to the local testing problem discussed before (e.g. the case of a weak signal in a signal detection model).

In general, the quantity $\frac{n_2}{n}$ is a function of α and $\{\beta_n\}_{n=1}^{\infty}$. However, under mild assumption the ARE is not dependent of these quantities. In particular, when

$$\tilde{\delta}_j^{(n)}(\underline{y}) = \begin{cases} 1 & \text{if } T_j^{(n)}(y_1, \cdots, y_n) > \tau_j^{(n)} \\ \gamma_j & \quad = \\ 0 & \quad < \end{cases}$$

and that the hypotheses are of the form

$$H_0: \ Y \sim P_{\theta_0}$$
$$H_1: \ Y \sim P_{\theta_n}$$

where $\theta_n > \theta_0$ and $\{P_\theta;\ \theta \geq \theta_0\}$ is a family of distributions for Y. In this case, $H_1^{(n)} \to H_0$ as $n\to\infty$ can be represented by $\lim_{n\to\infty} \theta_n = \theta_0$.

Define for $j=0,1$, $n=1,2,\cdots$, $\theta \geq \theta_0$:

$$\psi_j^{(n)}(\theta) = E\{T_j^{(n)}(Y_1, \cdots, Y_n) \mid Y \sim P_\theta\},$$

$$\sigma_j^{(n)}(\theta) = [Var(T_j^{(n)}(Y_1, \cdots, Y_n) \mid Y \sim P_\theta)]^{\frac{1}{2}},$$

i.e., the mean and standard deviation of the test statistic $T_j(Y)$ when $Y \sim P_\theta$, respectively.

Consider the following regularity conditions:

1. There exists a positive integer m such that the first through $(m-1)$th derivatives of $\psi_j^{(n)}(\theta)$ are zero at $\theta = \theta_0$, and
$$\frac{d^m}{d\theta^m} \psi_j^{(n)}(\theta)\Big|_{\theta=\theta_0} > 0 \quad \text{for } j = 0, 1.$$

2. There exists $\delta > 0$ such that, for $j = 0, 1$,
$$\lim_{n\to\infty} \Big[n^{-m\delta} \frac{d^m}{d\theta^m} \psi_j^{(n)}(\theta)\Big|_{\theta=\theta_0} \sigma_j^{(n)}(\theta_0) \Big] \triangleq c_j > 0.$$

3. Define $\theta_n = \theta_0 + K n^{-\delta}$ for $n = 1, 2, \cdots$. Then
$$\lim_{n\to\infty} \Big[\frac{d^m}{d\theta^m} \psi_j^{(n)}(\theta)\Big|_{\theta=\theta_n} \Big/ \frac{d^m}{d\theta^m} \psi_j^{(n)}(\theta)\Big|_{\theta=\theta_0} \Big] = 1$$
and $\lim_{n\to\infty} [\sigma_j^{(n)}(\theta_n) / \sigma_j^{(n)}(\theta_0)] = 1$.

4. Define $W_j^{(n)}(Y) = \Big[T_j^{(n)}(Y_1, \cdots, Y_n) - \psi_j^{(n)}(\theta)\Big] \Big/ \sigma_j^{(n)}(\theta)$.
Then
$$\lim_{n\to\infty} P_\theta\big(W_j^{(n)}(Y) \le w\big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{-x^2/2}\, dx \triangleq \Phi(w),$$
for all $w \in \mathbb{R}$, uniformly in θ for $\theta_0 \le \theta \le \theta_0 + d$ for some $d > 0$.

Then,

* Prop. III.C.1: The Pitman–Noether Theorem

Suppose that $\{\tilde{\delta}_1^{(n)}\}_{n=1}^{\infty}$ and $\{\tilde{\delta}_2^{(n)}\}_{n=1}^{\infty}$ satisfy Conditions 1–4, then, for the sequence of alternatives $\theta_n = \theta_0 + K n^{-\delta}$, we have

$$ARE_{1,2} = \eta_1/\eta_2$$

where η_j is defined by $\eta_j = (c_j)^{1/(m\delta)}$, $j=0,1$, and c_j is from Condition 2.

Remark 1. The quantity η_j is known as the (limiting) efficacy of the test sequence $\{\tilde{\delta}_j^{(n)}\}_{j=1}^{\infty}$. Thus the Pitman–Noether theorem asserts that the test sequence with higher efficacy is the most efficient asymptotically.

2. The regularity Conditions 1–4 are easily satisfied by many signal detection models. For example, consider the case in which the observations $Y_1, Y_2, \cdots$ are i.i.d. with marginal density $f_\theta(y)$. Consider

$$T_j^{(n)}(y_1, \cdots, y_n) = \sum_{k=1}^{n} g_j(y_k), \quad j=0,1.$$

Then $\psi_j^{(n)}(\theta) = n \int g_j f_\theta \, d\mu$

$$\sigma_j^{(n)}(\theta) = \left[n \left\{ \int (g_j)^2 f_\theta \, d\mu - (\psi_j^{(n)}(\theta))^2 \right\} \right]^{\frac{1}{2}}$$

Assume $\int g_j f_{\theta_0} d\mu = 0$ (0 mean)

and $\left. \dfrac{\partial \int g_j f_\theta d\mu}{\partial \theta} \right|_{\theta=\theta_0} > 0$

$\Rightarrow$ $m=1$ and $\delta = \frac{1}{2}$ and

$\eta_j = \left[\frac{\partial}{\partial \theta} \int g_j f_\theta d\mu \right]^2 \Big|_{\theta=\theta_0} \Big/ \int g_j^2 f_{\theta_0} d\mu.$

Exercises

5. Quaternary Phase-Shift Keying (QPSK) is an example of the situation in Exercise 3 with four signals ($M = 4$) given by

$$s_{lk} = E_0 \sin(\omega_c T(k-1) + (l+1/2)\pi/2), \quad k = 1, \ldots, n, \quad l = 0, \ldots, 3.$$

Assuming ω_c, T, and n are as in Example III.B.5, find the minimum error probability for equally likely signals in i.i.d. $\mathcal{N}(0, \sigma^2)$ noise. (Note that these signals are not orthogonal).

6. Suppose $\underline{Y} \sim \mathcal{N}(\underline{\mu}, \Sigma)$. For each $k \geq 2$, define $\hat{Y}_k = E\{Y_k | Y_1, \ldots, Y_{k-1}\}$ and $\hat{\sigma}^2_{Y_k} = \mathrm{Var}(Y_k | Y_1, \ldots, Y_{k-1})$. Also define $\hat{Y}_1 = E\{Y_1\}$ and $\hat{\sigma}^2_{Y_1} = \mathrm{Var}(Y_1)$. Define a sequence $I_1, I_2, \ldots, I_n$ by

$$I_k = (Y_k - \hat{Y}_k)/\hat{\sigma}_{Y_k}.$$

Show that $\underline{I} \sim \mathcal{N}(0, \mathbf{I})$, and thus that the above scheme provides whitening of $\underline{Y}$.

7. Consider the hypothesis pair

$$H_0 : Y_k = N_k, \qquad k = 1, \ldots, n$$

versus

$$H_1 : Y_k = N_k + \Theta S_k, \quad k = 1, \ldots, n$$

where $\underline{N} \sim \mathcal{N}(\underline{0}, \Sigma)$, $\underline{s}$ is known, and Θ is a random variable independent of $\underline{N}$.

(a) Find the α-level Neyman-Pearson detector and ROCs assuming that Θ is a discrete random variable taking the values $+1$ and -1 with equal probabilities (i.e., $P(\Theta = +1) = P(\Theta = -1) = 1/2$).

(b) Suppose that $\Theta \sim \mathcal{N}(0, \sigma_\theta^2)$. Assuming $\Sigma = \sigma^2 \mathbf{I}$, show that the likelihood ratio is of the form

$$L(\underline{y}) = k_1 e^{k_2 \|\underline{s}^T \underline{y}\|^2}$$

where k_1 and k_2 are positive constants. Find k_2.

§3. Sequential Detection

What wasstudied before are all fixed-sample size detectors, i.e., in each case, given a fixed number of observations, we want to derive an optimal detector based on these samples.

Sometimes, one may prefer to a fixed performance where the number of samples may vary. In other words, for some realizations of the observation sequence we may be able to make a decision after only a few samples, while for some other realizations, we may want to continue sampleing to make a better decision.

A detector that uses a random number of samples is known as a <u>sequential detector</u>. Mathematically, assume the observations Y_k, $k=1,2,\cdots$ are i.i.d.

$$H_0 : Y_k \sim P_0, \quad k=1,2,\cdots$$

vs

$$H_1 : Y_k \sim P_1, \quad k=1,2,\cdots$$

where P_0 and P_1 are two possible distributions. A <u>sequential decision rule</u> is a pair of sequences $(\underline{\phi}, \underline{\delta})$ where $\underline{\phi} = \{\phi_j;\ j=0,1,2,\cdots\}$ is called a stopping rule ($\phi_j : \mathbb{R}^j \to \{0,1\}$) and $\underline{\delta} = \{\delta_j;\ j=0,1,2,\cdots\}$ is called a terminal decision rule, δ_j being a decision rule on

$\{\mathbb{R}^j, \mathcal{B}^j\}$ for each j, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$.

The sequential decision $(\underline{\Phi}, \underline{\delta})$ operates as follows: For an observation sequence $\{y_k;\ k=1,2,\cdots\}$, the rule $(\underline{\Phi}, \underline{\delta})$ makes the decision $\delta_N(y_1, \cdots, y_N)$, where N is the stopping time defined by

$$N = \min\{n \mid \phi_n(y_1, \cdots, y_n) = 1\}.$$

That is, $\underline{\Phi}$ tells us when to stop taking samples by the mechanism that when $\phi_n(y_1, \cdots, y_n) = 0$, we take another sample [the $(n+1)$st] and when $\phi_n(y_1, \cdots, y_n) = 1$, we stop sampling and make a decision.

The number of samples, N, is random, since it depends on the data sequence.

The terminal decision rule $\underline{\delta}$ tells us what decision to make when we stop sampling.

An ordinary fixed-sample-size decision rule δ operating with n samples corresponds to

$$\phi_j(y_1, \cdots, y_n) = \begin{cases} 0 & \text{if } j \neq n \\ 1 & \text{if } j = n \end{cases}$$

$$\delta_j(y_1, \cdots, y_n) = \begin{cases} \delta(y_1, \cdots, y_n) & \text{if } j = n \\ \text{arbitrary} & \text{if } j \neq n. \end{cases}$$

To derive optimum sequential decision rules, we first consider the Bayesian decision, in which π_1 and $\pi_0 = 1-\pi_1$ are assigned to the hypotheses H_1 and H_0, respectively, and costs C_{ij} are assigned. For simplicity, we assume the uniform costs. In order to make the problem realistic, we should also assign a cost to observation. Thus, we assign a cost $C>0$ to each sample we take so that the cost of taking n samples is nC.

The conditional risks for a given sequential decision rule are

$$R_0(\underline{\phi}, \underline{\delta}) = E_0\{\delta_N(Y_1, \cdots, Y_N)\} + C\, E_0\{N\}$$

and

$$R_1(\underline{\phi}, \underline{\delta}) = 1 - E_1\{\delta_N(Y_1, \cdots, Y_N)\} + C\, E_1\{N\}$$

The Bayes risk is thus given by

$$r(\underline{\phi}, \underline{\delta}) = (1-\pi_1) R_0(\underline{\phi}, \underline{\delta}) + \pi_1 R_1(\underline{\phi}, \underline{\delta})$$

and a Bayesian sequential rule is one that minimizes $r(\underline{\phi}, \underline{\delta})$.

Let $$V^*(\pi_1) \triangleq \min_{\substack{\underline{\phi}, \underline{\delta} \\ \phi_0 = 0}} r(\underline{\phi}, \underline{\delta}), \quad 0 \le \pi_1 \le 1.$$

Since $\phi_0 = 0$ means that the test does not stop

With zero observations, $V^*(\pi_1)$ describes the minimum Bayes risk overall sequential tests that take at least one sample.

It is easy to show that $V^*(\pi_1)$ is a concave, continuous function of π_1 with $V^*(0) = V^*(1) = C$

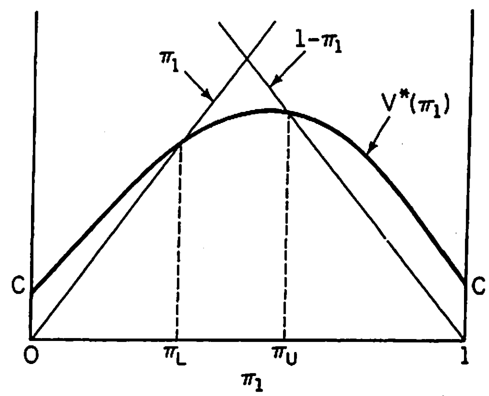

FIGURE III.D.1. Relationships yielding the Bayes sequential rule for uniform costs of errors and cost C per sample.

Also shown in the above figure are plots of the Bayes risk vs. π_1 for two other sequential decision rules: The one that takes no samples and decides H_1 (i.e., $\phi_0 = \delta_1 = 1$) and the one that takes no samples and decides H_0 (i.e., $\phi_0 = 1 - \delta_0 = 1$).

$$r(\underline{\phi}, \underline{\delta})\Big|_{\phi_0 = \delta_0 = 1} = 1 - \pi_1; \quad r(\underline{\phi}, \underline{\delta})\Big|_{\phi_0 = 1 - \delta_0 = 1} = \pi_1.$$

These two decision rules represent the only possible Bayes rules that are not included in the minimization. (Randomization cannot help with $\phi_0 = 1$, since if we choose $\delta_0 = 1$ with probability γ

and $\delta_0 = 0$ with probability $1-\gamma$, we get a Bayes risk $\gamma(1-\pi_1)+(1-\gamma)\pi_1$ which is always larger than $\min\{\pi_1, 1-\pi_1\}$.

By inspection of the above figure, we see that the Bayes rule for a fixed prior π_1 is $\phi_1 = 1-\delta_0 = 1$ if $\pi_1 \le \pi_L$, it is $\phi_0 = \delta_{10} = 1$ if $\pi_1 \ge \pi_U$, and it is the decision rule with minimum Bayes risk among all $(\underline{\Phi}, \underline{\delta})$ with $\phi_0 = 0$ if $\pi_L < \pi_1 < \pi_U$.

So, if $\pi_1 \le \pi_L$ we take no samples and choose H_0,
if $\pi_1 \ge \pi_U$ ... - - — — - - H_1.

Otherwise we take at least one sample.

Now, suppose that we have the condition

$$\pi_L < \pi_1 < \pi_U .$$

Here we know that the optimum test takes at least one sample. However, note that after having taken one sample, the problem of optimizing the test is conditionally the same as that with no samples, in the sense that we still have infinitely many i.i.d. samples at our disposal and the costs are the same. The one difference is that we now have taken a sample and so we have more information about which hypothesis is true. In particular, instead of having a prior probability π_1 we now have prior $\pi_1(y_1)$ that is actually the posterior probability of H_1 given our observation of Y_1, i.e.,

$$\pi_1(y_1) = P(H_1 \text{ is true} \mid Y_1 = y_1).$$

Thus, the picture after having taken one sample is exactly the same as the previous figure <u>except</u> that the abscissa variable π_1 is replaced with $\pi_1(y_1)$. Because the samples are independent, knowledge of Y_1 does not affect the shape of V^* (which now represents minimum risk over all tests that takes at least two samples). So, we conclude that after taking one sample the optimum test stops and choose H_0 if $\pi_1(y_1) \le \pi_L$, it stops and chooses H_1 if $\pi_1(y_1) \ge \pi_U$, and it takes another sample if $\pi_L < \pi_1(y_1) < \pi_U$.

If both $\pi_L < \pi_1 < \pi_U$ and $\pi_L < \pi_1(y_1) < \pi_U$, then the optimum test takes at least two samples. In this case we start over with the new prior $\pi_1(y_1, y_2) = P(H_1 \text{ is true} \mid Y_1 = y_1, Y_2 = y_2)$, and make the same comparison again.

Continuing this reasoning for an arbitrary number of samples taken, we see that the Bayes sequential test continues sampling until the quantity $\pi_1(y_1, \cdots, y_n) \triangleq P(H_1 \text{ is true} \mid Y_1 = y_1, \cdots, Y_n = y_n)$ falls out of the interval (π_L, π_U), and then it

Chooses H_0 if $\pi_1(y_1,\cdots,y_n) \le \pi_L$ and H_1 if $\pi_1(y_1,\cdots,y_n) \ge \pi_U$. This test is described by the stoping rule

$$\phi_n(y_1,\cdots,y_n) = \begin{cases} 0 & \text{if } \pi_L < \pi_1(y_1,\cdots,y_n) < \pi_U \\ 1 & \text{otherwise} \end{cases}$$

and the terminal decision rule

$$\delta_n(y_1,\cdots,y_n) = \begin{cases} 1, & \text{if } \pi_1(y_1,\cdots,y_n) \ge \pi_U \\ 0, & \quad\quad\quad\quad\quad\quad\quad\quad \le \pi_L \end{cases}$$

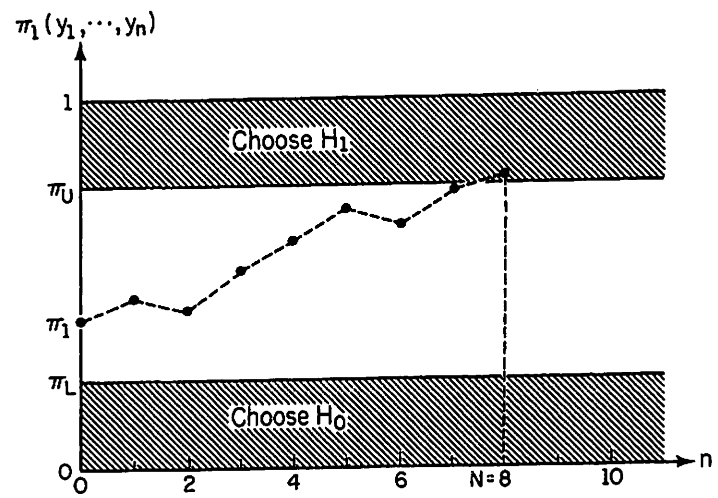

FIGURE III.D.2. Depiction of a realization of a Bayes sequential test.

Under mild conditions, $\pi_1(Y_1,\cdots,Y_n)$ converges almost surely to 1 under H_1 and to 0 under H_0. Thus the test terminates with probability 1.

All that is needed to specify the optimum test are the two probabilities π_L and π_U and a scheme to compute $\pi_1(y_1,\cdots,y_n)$. Unfortunately, it is difficult to obtain π_L and π_U exactly except in some special cases. On the other hand, to compute $\pi_1(y_1, \dots, y_n)$ is easy, in particular assuming that P_0 and P_1 have densities p_0 and p_1:

$$\pi_1(y_1, \dots, y_n) = \frac{\pi_1 \prod_{k=1}^{n} p_1(y_k)}{\pi_0 \prod_{k=1}^{n} p_0(y_n) + \pi_1 \prod_{k=1}^{n} p_1(y_k)} \quad \text{(III.D.8)}$$

$$= \frac{\pi_1 \lambda_n(y_1, \dots, y_n)}{\pi_0 + \pi_1 \lambda_n(y_1, \dots, y_n)}$$

where λ_n is the likelihood ratio based on n samples:

$$\lambda_n(y_1, \dots, y_n) = \prod_{k=1}^{n} \frac{p_1(y_k)}{p_0(y_k)}.$$

For consistency, we define $\lambda_0 = 1$.

$\pi_1(y_1, \dots, y_n)$ is monotonically increasing in λ_n. Thus, the test can be rewritten

$$\phi_n(y_1, \dots, y_n) = \begin{cases} 0 & \text{if } \underline{\pi} < \lambda_n(y_1, \dots, y_n) < \overline{\pi} \\ 1 & \text{otherwise} \end{cases} \quad (*)$$

and $$\delta_n(y_1, \dots, y_n) = \begin{cases} 1 & \text{if } \lambda_n(y_1, \dots, y_n) \geq \overline{\pi}, \\ 0 & \quad " \quad\quad\quad\quad \leq \overline{\pi}, \end{cases}$$

where $\underline{\pi} \triangleq \frac{\pi_0 \pi_L}{\pi_1(1 - \pi_L)}$, $\overline{\pi} \triangleq \frac{\pi_0 \pi_U}{\pi_1(1 - \pi_U)}$.

Thus, the Bayes sequential test takes

samples until the likelihood ratio falls outside the interval $(\underline{\pi}, \overline{\pi})$ and then it decides on H_0 or H_1 depending on whether λ_n falls below $\underline{\pi}$ or above $\overline{\pi}$.

(*) is an example of a sequential probability ratio test (SPRT). In particular, for any real numbers A and B, denoted by SPRT(A, B), is defined as in (*) with $\underline{\pi}$ replaced by A and $\overline{\pi}$ replaced by B, and until the decision rule left arbitrary if $A = B$. Thus, the SPRT(A,B) continues sampling until the likelihood ratio λ_n falls outside the "boundaries" A and B and then chooses H_1 if $\lambda_n \geq B$ and H_0 if $\lambda_n \leq A$. Note that if $A = 1 < B$, we take no samples and choose H_0; if $A < B = 1$, we take no samples and choose H_1, and if $A = B = 1$, we take no samples and take an arbitrary choice.

* Example : Sequential Detection of a Constant Signal

Consider the problem of detecting a constant signal in additive i.i.d. noise:

$$H_0 : Y_k = N_k, \quad k = 1, 2, \cdots$$

vs

$$H_1 : Y_k = N_k + \theta, \quad k = 1, 2, \cdots$$

where $\theta > 0$ and $\{N_k\}_{k=1}^{\infty}$ is an i.i.d. sequence of $\mathcal{N}(0, \sigma^2)$ noise samples. Then,

$$\lambda_n(y_1, \cdots, y_n) = \exp\left\{ \frac{\theta}{\sigma^2} \sum_{k=1}^{n} \left(y_k - \frac{\theta}{2} \right) \right\}$$

$\Rightarrow$ SPRT(A, B) computes $\frac{\theta}{\sigma^2} \sum_{k=1}^{n} \left(y_k - \frac{\theta}{2} \right)$

at each stage n and compare it with $\log A$ and $\log B$, stopping and choosing H_1 when the quantity exceeds $\log B$, and H_0 when the quantity falls below $\log A$.

* SPRT(A, B) has another property called the Wald-Wolfowitz Theorem (which is analogous to the Neyman-Pearson Lemma).

For a sequential decision rule $(\underline{\phi}, \underline{\delta})$, let $P_F(\underline{\phi}, \underline{\delta})$ denote the probability of a false alarm and let $P_M(\underline{\phi}, \underline{\delta})$ denote the probability of a miss, i.e.,

$$P_F(\underline{\phi}, \underline{\delta}) = P(\delta_N(Y_1, \cdots, Y_N) = 1 \mid H_0)$$

$$P_M(\underline{\phi}, \underline{\delta}) = P(\delta_N(Y_1, \cdots, Y_N) = 0 \mid H_1)$$

Also, let $N(\underline{\phi})$ denote the random stopping time associated with $\underline{\phi}$, i.e.,

$$N(\underline{\phi}) = \min\{ n \mid \phi_n(Y_1, \cdots, Y_n) = 1 \},$$

also known as the sample number of $\underline{\phi}$.

Proposition III.D.1: The Wald–Wolfowitz Theorem:

Suppose that $(\underline{\phi}_0, \underline{\delta}_0)$ is the SPRT(A, B) and that $(\underline{\phi}, \underline{\delta})$ is any other sequential decision rule for which

$$P_F(\underline{\phi}, \underline{\delta}) \le P_F(\underline{\phi}_0, \underline{\delta}_0)$$

and $$P_M(\underline{\phi}, \underline{\delta}) \le P_M(\underline{\phi}_0, \underline{\delta}_0).$$

Then, $$E\{N(\underline{\phi}) | H_j\} \ge E\{N(\underline{\phi}_0) | H_j\} \text{ for } j=0,1.$$

(It follows from the Bayes optimality).

⇒ For a given level of performance, no sequential decision rule has a smaller expected sample size than does the SPRT with that performance.

⇒ The average sample size of an SPRT is no larger than the sample size of a fixed-sample-size test with the same performance.

For a given expected sample size, no sequential decision rule has smaller error probabilities than does the SPRT.

* SPRT(A, B) allows us to choose P_F and P_M arbitrarily at the expense of larger expected sample size as below.

Suppose that $(\underline{\phi}, \underline{\delta})$ is the SPRT(A, B) with $A < 1 \le B$, and let $\alpha = P_F(\underline{\phi}, \underline{\delta})$, $\gamma = (1-\beta) = P_M(\underline{\phi}, \underline{\delta})$, $N = N(\underline{\phi})$. The rejection region of $(\underline{\phi}, \underline{\delta})$ can be written as $\Gamma_1 = \{ y \in \mathbb{R}^\infty \,/\, \lambda_N(y_1, \cdots, y_N) \ge B \} = \bigcup_{n=1}^{\infty} Q_n$

where $Q_n = \{ y \in \mathbb{R}^\infty \,|\, N = n \text{ and } \lambda_n(y_1, \cdots, y_n) \ge B \}$.

It is not hard to see $Q_n \cap Q_m = \phi$ for $m \ne n$.

$$\alpha = P(\lambda_N(Y_1, \cdots, Y_N) \ge B \,|\, H_0)$$

$$= \sum_{n=1}^{\infty} \int_{Q_n} \prod_{k=1}^{n} [p_0(y_k)\, \mu(dy_k)]$$

On Q_n, we have $\prod_{k=1}^n p_0(y_k) \le B^{-1} \prod_{k=1}^n p_1(y_k)$,

so that $\alpha \le B^{-1} \sum_{k=1}^{\infty} \int_{Q_n} \prod_{k=1}^{n} [p_1(y_k)\, \mu(dy_k)]$

$$= B^{-1} P(\lambda_N(Y_1, \cdots, Y_N) \ge B \,|\, H_1)$$

$$= B^{-1}(1-\gamma)$$

Similarly

$$\gamma = P(\lambda_N(Y_1, \cdots, Y_N) \le A \,|\, H_1)$$

$$\le A\, P(\lambda_N(Y_1, \cdots, Y_N) \le A \,|\, H_0)$$

$$= A(1-\alpha)$$

Thus, we have $B \le (1-\gamma)/\alpha$ and $A \ge \gamma/(1-\alpha)$.

(III. D. 14)

⇒ We can use the inequalities (III.D.14) to get approximate values for boundaries A and B to give desired α and γ, by assuming that when the likelihood ratio λ_n crosses a boundary, the excess over the boundary, i.e., $\lambda_N(Y_1, \dots, Y_N) - B$ or $A - \lambda_N(Y_1, \dots, Y_N)$, is negligible. This approximation will be accurate if N is relatively large on the average. Thus, we assume that either

$$\lambda_N(Y_1, \dots, Y_N) \cong A \quad \text{or} \quad \lambda_N(Y_1, \dots, Y_N) \cong B$$

and (III.D.14) become approximate equalities:

$$B \cong \frac{1-\gamma}{\alpha} \quad \text{and} \quad A \cong \frac{\gamma}{1-\alpha} . \qquad \text{(III.D.15)}$$

These approximations are known as Wald's approximations.

Suppose that α_d and γ_d are desired error probabilities and that we use the approximations (III.D.15) to choose the actual boundaries, i.e.,

$$A_a = \frac{\gamma_d}{1-\alpha_d}, \quad B_a = \frac{1-\gamma_d}{\alpha_d} .$$

Then, the actual error probabilities α_a and γ_a will satisfy (III.D.14):

$$\alpha_a/(1-\gamma_a) \le B_a^{-1} = \alpha_d/(1-\gamma_d)$$

$$\gamma_a/(1-\alpha_a) \le A_a \le \gamma_d/(1-\alpha_d)$$

$$\Rightarrow \alpha_a \le \frac{\alpha_d(1-\gamma_a)}{1-\gamma_d} \le \alpha_d/(1-\gamma_d)$$
(III.D.17)

$$\gamma_a \leq \frac{\gamma_d(1-\alpha_a)}{1-\alpha_d} \leq \frac{\gamma_d}{1-\alpha_d} \qquad \text{(III.D.17)}$$

So, for example, if $\gamma_d = \alpha_d$, we have

$$\alpha_a \leq \alpha_d + O(\alpha_d^2)$$

$$\gamma_a \leq \gamma_d + O(\gamma_d^2)$$

The inequalities (III.D.17) guarantees that we can obtain arbitrarily good error prob. performance by proper choice of the boundaries A and B.

(III.D.17) is also exact, not based on any approximation.

These are irrelevant to the actual distribution of Y_k's, which is a significant advantage over a fixed sample-size test.

* How does one evaluate the expected sample size of a sequential detector? To answer this question, we first consider a slightly more general sequential test for H_0 and H_1 as follows:

For each $a<0<b$ and each function $g: \mathbb{R} \to \mathbb{R}$, define the sequential decision rule $ST(a,b;g)$ by the pair $(\underline{\Phi}, \underline{\delta})$ given

by

$$\phi_j(y_1, \ldots, y_j) = \begin{cases} 0 & \text{if } a < \sum_{k=1}^{j} g(y_k) < b \\ 1 & \text{otherwise,} \end{cases}$$

and

$$\delta_j(y_1, \ldots, y_j) = \begin{cases} 1 & \text{if } \sum_{k=1}^{j} g(y_k) \geq b \\ 0 & \text{if } \quad \cdot\cdot \quad \leq a \end{cases}$$

Note that for $0 < A < 1 < B < \infty$ and $\frac{p_1}{p_0} < \infty$, the SPRT(A, B) is $ST(a, b; g)$ with $a = \log A$, $b = \log B$, and $g = \log\left(\frac{p_1}{p_0}\right)$. For the case in Example, $g(x) = \theta(x - \frac{\theta}{2})/\sigma^2$.

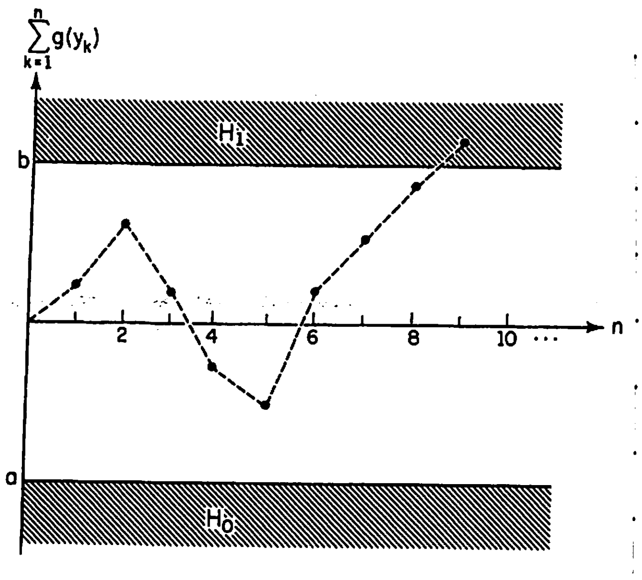

FIGURE III.D.3. Illustration of $ST(a, b; g)$.

* Proposition III.D.2 [Wald's identity] : The Fundamental Identity of Sequential Analysis

Suppose $(\underline{\phi}, \underline{\delta}) = ST(a, b; g)$. Define

$N = \min\{n \mid \phi_n(Y_1, \ldots, Y_n) = 1\}$ and

$$S_n = \sum_{k=1}^{n} g(Y_k)$$

Let M_j denote the moment-generating function of the random variable $g(Y_1)$ under hypothesis H_j, i.e.,

$$M_j(t) = E\{\exp\{t\, g(Y_1)\} | H_j\}, \quad j=0,1.$$

Suppose that $j=0$ or 1. Then, if $P(g(Y_1)=0|H_j) \neq 1$ and $P(|g(Y_1)| < \infty | H_j) = 1$; we have

$$E\{\exp\{t\, S_N\}[M_j(t)]^{-N} \;|H_j\} = 1$$

for all real t for which $M_j(t) < \infty$.

It is a consequence of the optional sampling theorem for martingales [Breiman 1968].

* Proposition III. D. 3 : Corollary to Wald's Identity

Under the hypothesis of the Wald's identity. Suppose that $M_j(t) < \infty$ in a neighborhood of $t=0$. Define $\mu_j = E\{g(Y_1) | H_j\}$ and $\sigma_j^2 = Var(g(Y_1) | H_j)$. Then

(a) $E\{S_N | H_j\} = \mu_j \, E\{N | H_j\}$

(b) $E\{(S_N - N\mu_j)^2 | H_j\} = \sigma_j^2 \, E\{N | H_j\}$.

* Extend Wald's approximation to the test $ST(a,b;g)$

Suppose that we can find two nonzero numbers t_0 and t_1 such that $M_j(t_j)=1$ for $j=0,1$. (the existence of such a t_j is assured if $P_j(g(Y_1)<0)>0$, $P_j(g(Y_1)>0)>0$, and $\mu_j \neq 0$ see Ferguson 1967). Then, Wald's identity implies

$$E\{\exp\{t_j S_N\} \mid H_j\} = 1 \text{ for } j=0,1. \qquad (\text{III.D.18})$$

If we ignore the "excess over the boundaries," then under H_0, S_N is a discrete random variable taking the values b and a with probabilities $P_F(\underline{\phi},\underline{\delta}) \triangleq \alpha$ and $(1-\alpha)$, respectively, and under H_1, S_N takes values a and b with probabilities $\gamma = P_M(\underline{\phi},\underline{\delta})$ and $1-\gamma$, respectively. Thus, (III.D.18) implies

$$(1-\alpha)e^{t_0 a} + \alpha e^{t_0 b} \cong 1$$

and $$\gamma e^{t_1 a} + (1-\gamma)e^{t_1 b} \cong 1.$$

$$\Rightarrow \alpha \cong (1-e^{t_0 a})/(e^{t_0 b} - e^{t_0 a})$$

$$\gamma \cong (1-e^{t_1 b})/(e^{t_1 a} - e^{t_1 b})$$

As an example, suppose that we take

$$g = \log\left(\frac{p_1}{p_0}\right)$$

So that $$ST(a,b;g) = SPRT(e^a, e^b).$$

Then, $$M_0(t) = \int_{\mathbb{R}} \exp\{t \log[p_1(y)/p_0(y)]\}\, p_0(y)\,\mu(dy)$$
$$= \int_{\mathbb{R}} [p_1(y)]^t [p_0(y)]^{1-t}\,\mu(dy)$$

Notice that $M_0(1) = \int_{\mathbb{R}} p_1(y)\,\mu(dy) = 1.$

$\Rightarrow \quad t_0 = 1.$

Similarly, $$M_1(t) = \int_{\mathbb{R}} [p_1(y)]^{t+1} [p_0(y)]^{-t}\,\mu(dy)$$

$\Rightarrow \quad t_1 = -1.$

Let $A = e^a,\ B = e^b.$

$\Rightarrow \quad \alpha \cong \dfrac{1-A}{1-B}, \quad \gamma \cong \dfrac{1-B^{-1}}{A^{-1}-B^{-1}}$

which are equivalent to the Wald's approximations in (III.D.15).

* If $\mu_0 = 0 \ (= E\{g(Y_1)|H_0\})$, from the corollary of Wald's identity, we have
$$E\{S_N | H_0\} = 0.$$

Thus, $$a(1-\alpha) + b\alpha \cong 0$$

$$\text{or } \alpha \cong \frac{-a}{b-a}.$$

Similarly, if $\mu_1 = 0$, $\gamma a + (1-\gamma) b \cong 0$.

$$\Rightarrow \quad \gamma \cong \frac{-b}{a-b}.$$

$\Rightarrow$ If $\mu_0 \neq 0$, by ignoring the excess over the boundaries, the corollary implies

$$E\{N|H_0\} = \frac{1}{\mu_0} E\{S_N|H_0\} \cong \frac{1}{\mu_0}(a(1-\alpha) + b\alpha)$$

If $\mu_0 = 0$, we use (b)

$$E\{N|H_0\} = \frac{1}{\sigma_0^2} E\{S_N^2|H_0\} \cong \frac{1}{\sigma_0^2}[a^2(1-\alpha) + b^2\alpha]$$

Similar expressions hold for $E\{N|H_1\}$:

If $\mu_1 \neq 0$,

$$E\{N|H_1\} = \frac{1}{\mu_1} E\{S_N|H_1\} \cong \frac{1}{\mu_1}(a\gamma + b(1-\gamma))$$

If $\mu_1 = 0$,

$$E\{N|H_1\} = \frac{1}{\sigma_1^2} E\{S_N^2|H_1\} \cong \frac{1}{\sigma_1^2}[a^2\gamma + b^2(1-\gamma)]$$

* For the SPRT(e^a, e^b), we have $\mu_j \neq 0$ for $j = 0, 1$, and Wald's approximations yield

$$E\{N|H_0\} \cong \frac{1}{\mu_0}\left[(1-\alpha)\log\frac{\gamma}{1-\alpha} + \alpha\log\frac{1-\gamma}{\alpha}\right]$$

$$E\{N|H_1\} \cong \frac{1}{\mu_1}\left[\gamma\log\frac{\gamma}{1-\alpha} + (1-\gamma)\frac{1-\gamma}{\alpha}\right]$$

(III. D. 23)

* Example 2 : A Comparison of Sequential and FSS Detectors

Fixed-sample-size

The best FSS detector is the likelihood ratio detector, which for a given sample size n, has error probabilities α and γ related through the expression (III. B.31)

$$(1-\gamma) = 1 - \Phi\left(\Phi^{-1}(1-\alpha) + n^{1/2}\,\theta/\sigma\right)$$

$$\Rightarrow \quad n_{FSS} = \left\lceil \sigma^2 \left[\Phi^{-1}(1-\alpha) - \Phi^{-1}(\gamma)\right]^2 / \theta^2 \right\rceil$$

where $\lceil x \rceil$ denotes the smallest integer that is not smaller than x.

For the SPRT with boundaries chosen for given error probabilities, the expected sample sizes are given by (III. D.23). To evaluate them, we must compute μ_j :

$$\mu_j = E\{\log[p_1(Y_1)/p_0(Y_1)] \mid H_j\}$$

$$= \frac{\theta}{\sigma^2} E\{(Y_1 - \tfrac{\theta}{2}) \mid H_j\}$$

$$= \begin{cases} -\dfrac{\theta^2}{2\sigma^2} & \text{if } j=0 \\ \dfrac{\theta^2}{2\sigma^2} & \text{if } j=1. \end{cases}$$

For simplicity, we assume that $\alpha = \gamma$. Then

$$E\{N|H_1\} = E\{N|H_0\} \cong 2\sigma^2 \left[\gamma \log \frac{\gamma}{1-\alpha} + (1-\gamma)\log\frac{1-\gamma}{\alpha}\right]$$

For example, $\alpha = \gamma = 0.1$ and $\frac{\theta^2}{\sigma^2} = 1$.
Then, $N_{FSS} \cong 22$ and $E\{N|H_j\} \cong 9$.

$$\lim_{\alpha=\gamma\to 0} \frac{E\{N|H_j\}}{N_{FSS}} = \frac{1}{4}.$$

* Example III.D.2 illustrates that the SPRT can, on the average, offer substantial savings over the best FSS test in terms of the numbers of samples required to perform a test with a given level of performance. This is particularly advantageous in applications in which a large number of identical tests are to be performed. An example of such an application is search radar in which the radar performs a test (target present vs. target absent) in each of many cells in a search area.

* Why not abandon the use of FSS likelihood ratio tests in favor of SPRTs in all cases?

Several practical disadvantages of SPRTs.

Disadvantage 1: Although the sample size of an SPRT is finite with probability 1, it is not bounded.

The SPRT saves samples by making quick decisions when the hypothesis is clear from observed data; but on the other hand, if the observed data are ambiguous, the SPRT can run on for a large number of samples.

As the Wald-Wolfowitz theorem implies the average of these two effects is benefical, however, the occasional long run may not be practical for many applications.

This difficulty can be overcome quite easily by modifying the SPRT to stop sampling and make a hard (single-threshold) decision after some maximum number of samples, called truncated SPRT.

Disadvantage 2: Their implementation requires an exact knowledge of both p_0 and p_1. For example in the previous Example 1, it is necessary to know the signal value θ to implement the test. This is in contrast to the FSS Neyman-Person test for the same problem, which is uniformly most powerful for $\theta > 0$.

Disadvantage 3: The theory is limited when the i.i.d. assumption can not be involved.

23. Consider the problem of detecting a $\mathcal{N}(\underline{0}, \Sigma_S)$ signal in $\mathcal{N}(\underline{0}, \sigma^2 \mathrm{I})$ noise with $n = 2$ and

$$\Sigma_S = \sigma_S^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

For equally likely priors compute and compare the exact error probability and the Chernoff bound on the error probability for $\rho = 0.0, \rho = -0.5$, and $\rho = +0.5$, and for $\sigma_S^2/\sigma^2 = 0.1, \sigma_S^2/\sigma^2 = 1.0$, and $\sigma_S^2/\sigma^2 = 10.0$.

24. Investigate the Chernoff bound for testing between the two marginal densities

$$p_0(y) = \begin{cases} 1 & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_1(y) = \begin{cases} 2y & \text{if } 0 \le y \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

for a sequence of i.i.d. observations, $Y_1, Y_2, \ldots, Y_n$.

25. Consider a sequence of i.i.d. Bernoulli observations, $Y_1, Y_2, \ldots$, with distribution

$$P(Y_k = 1) = 1 - P(Y_k = 0) = 1/3$$

under hypothesis H_0, and

$$P(Y_k = 1) = 1 - P(Y_k = 0) = 2/3$$

under hypothesis H_1.

(a) Use Wald's approximations to suggest values of A and B so that the SPRT (A, B) has maximum error probability $p^* = \max(P_F, P_M)$ approximately equal to 0.01. Describe the resulting test in detail. Also, using Wald's approximations, give an approximation to the expected sample sizes $E\{N|H_0\}$ and $E\{N|H_1\}$.

(b) Find an integer n as small as you can so that the maximum error probability for the optimal test with fixed sample size n is no more than 0.01. Compare n to the expected sample sizes found in part (a) (*Note*: You may use a Chernoff bound to find n, rather than finding the smallest possible n.)

(c) Compute $p^*, E\{N|H_0\}$ and $E\{N|H_1\}$ exactly for the test you found in part (a), and compare with the approximate values you found in part (a). [*Hint*: Use the fact that the SPRT you found in part (a) is equivalent to SPRT (A', B') where A' and B' are integer powers of 2.]