Detection and Estimation
Chapter 3. Topics in Signal Detection

Husheng Li

Min Kao Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

Spring, 2015
A detector that uses a random number of samples depending on the observation sequence is called sequential detection.

We assume that the observations \( \{ Y_t \} \) are i.i.d, and consider two hypotheses \( H_0 : Y_t \sim P_0 \) and \( H_1 : Y_t \sim P_1 \).

A sequential rule \( (\phi, \delta) \) where \( \phi \) is the stopping rule (when to stop receiving observations) and \( \delta \) is the terminal decision rule (what is the output hypothesis).
Cost Structure

- The conditional risk is given by

\[ R_j(\phi, \delta) = E_0[\delta_N(Y_1, ..., Y_N)] + CE_j[N], \]

where \( N \) is the stopping time. We assume uniform cost for the detection and cost \( C \) per observation.

- The Bayesian risk is then given by

\[ r(\phi, \delta) = (1 - \pi_1)R_0 + \pi_1 R_1. \]

- Consider the function which is concave in \( \pi_1 \):

\[ V^*(\pi_1) = \min_{\phi,\delta} r(\phi, \delta). \]
The two lines represent the decision rules of no observations. When $\pi_1 < \pi_L$, we take no samples and choose $H_0$. When $\pi_1 > \pi_U$, we take no samples and choose $H_1$. 
Bayesian Decision Rule

Bayesian Rule: We update the prior probability $\pi_1$, compare with $\pi_L$ and $\pi_U$ and then make the decision. This is an example of sequential probability ratio test (SPRT).

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

SPRT was proposed by A. Wald (1902-1950). He studied the problem of bomber losses to enemy fire (what is the most places in an aircraft that need the most reinforcement of armor?). He was attacked by Fisher after his death in an airplane crash.

The Wald-Wolfowitz Theorem: Compared with the $SPRT(A, B)$, if there is another decision rule that results in less false alarm and missed detection rates, then it must result in a larger expectation of observation number.
Suppose that we want to achieve false alarm and missed detection rates $\delta$ and $\gamma$.

Wald Approximations:

\[
\begin{align*}
B & \approx (1 - \gamma)/\alpha \\
A & \approx \gamma/(1 - \alpha)
\end{align*}
\]
Consider the decision rule $ST(a, b, g)$.

Wald identity:

$$E \left[ e^{tS_N} \frac{1}{M_j^N(t)} \big| H_j \right] = 1,$$

where $M_j$ is the generating function of $g(Y)$ under $H_1$.

Corollary: $E[S_N|H_j] = \mu_j E[N|H_j]$ and $E \left[ (S_N - N\mu_j)^2 | H_j \right] = \sigma^2_j E[N|H_j]$
Disadvantages of SPRT

- Although the sample size of an SPRT is finite with probability one, it is unbounded.
- The implementation requires perfect knowledge of $p_0$ and $p_1$.
- The theory of these tests is limited when the i.i.d. assumption is violated.
Nonparametric Detection

- All the previous detectors are based on the knowledge of distributions. But sometimes we only have approximate information about the distributions. Then, we have two approaches to handle the uncertainty in the distributions, namely nonparametric detection and robust detection.

- Consider the hypothesis testing $H_i : Y_k \sim P \in P_i, i = 0, 1$. If classes $P_0$ and $P_1$ are parametrized by a real or vector parameter.

- Example: $H_0 : Y_k = N_k$ and $H_1 : Y_k = N_k + \theta$. What if we only know that $N_k$ is symmetric about zero?
Principle of Nonparametric Detection

- Generally speaking, a nonparametric test is one designed to operate over wide classes $P_0$ and $P_1$ with some performance characteristic being invariant over the classes.
- The tests are usually designed to be simple, using rough information about the data (e.g., signs, ranks, et al), instead of the exact values of data.
- In some applications, the nonparametric tests are sometimes called constant-false-alarm (CFAR) detectors.
The Sign Test

- Assume i.i.d. observations \( \{Y_k\} \). Define \( p = P(Y_1 > 0) \). Consider the hypothesis pair

\[
H_0 : p = \frac{1}{2} \quad H_1 : \frac{1}{2} < p < 1.
\]

- The sign test is given by

\[
\delta(y) = \begin{cases} 
1, & > \\
\gamma, & \text{if } t(y) = \tau \\
0, & <
\end{cases}
\]

where \( t(y) = \sum_{k=1}^{n} u(y_k) \) and \( u \) is the step function.
The Rank Test

- We can use more information without the loss of the nonparametric characteristics.
- We can replace the sign test statistic with

\[ t(y) = \sum_{k=1}^{n} u(y_k) \rightarrow t(y) = \sum_{k=1}^{n} \lambda_k u(y_k), \]

where \( \lambda_k \) is the rank of \( y_k \) when \( \{y_j\} \) are reordered in increasing order.
- The resulting test is called the Wilcoxon test.
Robust Detection

- Suppose that the data is not exactly $P_0$ or $P_1$, say, the actual distribution is

$$(1 - \epsilon)P_j + \epsilon M_j, j = 0, 1,$$

where $\{M_j\}$ are ‘contaminating’ distributions.

- Then, $M_0$ could place all its probabilities in regions where $p_1(y_k) \ll p_0(y_k)$ such that more false alarms are generated (in the order of $1 - (1 - \epsilon)^n$).

- Anything can be done to alleviate the lack of robustness? Can we simply add a limiter to the likelihood ratio?
Huber’s Minimax Approach

- A reasonable design criterion is to replace the usual error probabilities with their worst case values.
- We assume that the true distributions lie in the neighborhood $\mathcal{P}_0$ of $P_0$ and in the neighborhood $\mathcal{P}_1$ of $P_1$. Then, we define

$$P_F(\delta, \mathcal{P}_0) = \sup_{P \in \mathcal{P}_0} P_F(\delta, P), \quad P_F(\delta, \mathcal{P}_1) = \sup_{P \in \mathcal{P}_1} P_M(\delta, P)$$

- For the model of $\epsilon$-contaminated model, we just replace $p_0$ and $p_1$ with

$$q_0(y_k) = \begin{cases} (1 - \epsilon)p_0(y_k), & \text{if } p_1(y_k) < c''p_0(y_k) \\ \frac{(1 - \epsilon)}{c''}p_1(y_k), & \text{if } p_1(y_k) \geq c''p_0(y_k) \end{cases}$$

$$q_0(y_k) = \begin{cases} (1 - \epsilon)p_1(y_k), & \text{if } p_1(y_k) > c'p_0(y_k) \\ c'(1 - \epsilon)p_0(y_k), & \text{if } p_1(y_k) \leq c'p_0(y_k) \end{cases}.$$
Quickest detection is to detect the change of distribution in a random process.

Consider a random process having independent variables. At the beginning the distribution of $X(t)$ is $P_0$. At an unknown time, it is changed to $P_1$. It is desired to detect the change as quickly as possible.
Optimal Stopping Time

- We can consider the detection of the change as selecting an optimal stopping time.

- Consider a random process with probability space \((\Omega, \mathcal{F}, P)\), a random time \(T\) is called a stopping time if \(\{T \leq t\} \in \mathcal{F}_t\).

- For a class \(S\) of stopping times, the payoff is

\[
V(S) = \sup_{T \in S} E[Y_T],
\]

where \(Y_t\) is the reward that can be claimed at time \(t\).
Finite-horizon Case

- We consider a time horizon with maximum time $n$. The family of stopping times is $S^n = \{ T \in \mathcal{T} | T \leq n \}$, where $\mathcal{T}$ is the set of all stopping times with respect to $\{\mathcal{F}_t\}$. We also define $S^n_k = \{ T \in \mathcal{T} | k \leq T \leq n \}$.
- Define $\gamma^n_k$ the optimal payoff if the stopping time is not before time $k$, namely
  \[
  \gamma^n_k = \text{esssup}_{T \in S^n_k} E[Y_T | \mathcal{F}_k].
  \]
  Then, the optimal stopping time is
  \[
  T^n = \inf\{ k \geq 0 | \gamma^n_k = Y_k \},
  \]
i.e., it stops at the first time $k$ such that the current reward $Y_k$ equals the largest conditional expected reward.
- $\gamma^n_k$ can be calculated using dynamics programming.
Suppose that the prior probability of the change time $t$ is known and is given by $P(t)$. Then, we define

$$\pi_k = P(t \leq k \leq k|\mathcal{F}_k), \quad k = 0, 1, ...$$

Consider the cost defined as

$$P(T < t) + cE \left[(T - t + 1)^+\right],$$

which is a combination of false alarm rate and detection delay.
We assume

\[ P(t = k) = \begin{cases} 
\pi, & \text{if } k = 1 \\
(1 - \pi)\rho(1 - \rho)^{k-1}, & \text{if } k = 1, 2, \ldots
\end{cases} \]

Then, the optimal time to claim the change is

\[ T_B = \{ k \geq 0 | \pi_k \geq \pi^* \}, \]

where \( \pi^* \) is a predetermined threshold.
Non-Bayesian Quickest Detection

Now we assume that the prior information about the change point is unknown. Then, we define

- Worst case detection delay:
  \[ d(T) = \sup_{t \geq 1} \text{esssup}_t [(T - t + 1)^+ | \mathcal{F}_{t-1}] \], where the worst case esssup is over all realization of random processes.

- Mean time between false alarms: \( f(T) = E_\infty[T] \)

Then, the design criterion is given by

\[ \inf_{T \in \mathcal{T}} d(T), \text{ s.t. } f(T) \geq \gamma. \]
CUSUM Stopping Time

- The CUSUM (cumulative sum) stopping time is
  
  \[ T^c_h = \inf\{k \geq 0 | S_k \geq h\}, \]

  where the metric \( S_k \) is given by

  \[ S_k = \max\{S_{k-1}, 1\} L(Z_k), \quad k \geq 1, \]

  where \( L(Z_k) \) is the likelihood ratio calculated from the \( k \)-th observation \( Z_k \).

- The metric \( S_k \) is also given by

  \[ S_k = \max_\tau \left\{ \prod_{t=\tau}^k L(Z_t) \right\}. \]

- The CUSUM algorithm is proposed in 1930s. Its optimality is proved in 1985 (by Moustakides).
Performance of CUSUM Test: Exact Result

- The expected false alarm and delay are given by

\[ f(T_h^c) = \frac{E_\infty[N]}{1 - P_\infty(F_0)}, \]

and

\[ d(T_h^c) = \frac{E_1[N]}{1 - P_1(F_0)}, \]

where \( N \) is the stopping time \( N = \min \left\{ n \geq \sum_{i=1}^{n} \log L(Z_i) \notin (0, \log h) \right\} \) and \( F_0 \) means the event \( \sum_{i=1}^{N} \log L(Z_i) \leq 0. \)
Performance of CUSUM Test: Lorden’s Approach

- Lorden relates the CUSUM test to the SPRT: Suppose $T$ is a stopping time with respect to $\sigma\{Z_1, ..., Z_k\}$ such that $P_\infty(T < \infty) \leq a$. Let $T_k$ be the stopping time obtained by applying $T$ to $Z_k, Z_{k+1}, ...$, and define

$$T^* = \inf\{T_k + k - 1 | k = 1, 2, ...\}.$$

Then, we have $E_\infty[T^*] \geq 1/2$ and $d(T^*) \leq E_1(T)$.

- Then, for the CUSUM test, we have

$$E_\infty T^* \geq \frac{1}{a}, \quad d(T^*) \leq \frac{|\log a|}{l_1},$$

where $l_1 = E_1[\log L(Z_1)]$ and $a = 1/h$. 


In the Shiryaev-Roberts Stopping time scheme, the metric is calculated in the following steps:

- Set $R_0 = 0$.
- Compute $R_{n+1} = L(Z_{n+1})(1 + R_n)$.
- If $R_n \geq h$, where $h$ is a threshold, claim the change.

In practice, the Shiryaev-Roberts scheme has a similar performance to that of CUSUM test.
Multiple Hypothesis Testing

- Suppose the size of data is $n$. A multiple hypothesis testing (MTP) provides rejection regions $C_n(m)$ to determine whether the corresponding null hypothesis $H_0(m)$ should be rejected; i.e., providing

$$R_n = \{ m : H_0(m) \text{ is rejected} \}.$$

- The notation $R(T_n, Q_{0m}, \alpha)$ and $C(m; T_n, Q_{0n}, \alpha)$ provides the following ingredients:
  - $T_n$: test statistics.
  - $Q_{0m}$: the null distribution of $T_n$ in the $m$-th hypothesis.
  - $\alpha$: the nominal Type I error level.
We consider MTPs based on nested rejection regions:

\[ C(m; T_n, Q_{0n}, \alpha_1) \subset C(m; T_n, Q_{0n}, \alpha_2), \text{ if } \alpha_1 \leq \alpha_2. \]

Different approaches:
- Marginal vs. joint MTP.
- Single-step vs. stepwise MTP.
- Common-cut-off vs. common quantile MTP.
Errors in MTP

- Type I (reps. II), or false positive (reps. false negative) error means rejecting a true null (reps. fails to reject a false null) hypothesis. The number of rejected null hypotheses is $R_n$ and the number of Type I errors is denoted by $V_n$.

- Different Type I Error Rates:
  - The family wise error rate: $FWER = Pr(V_n > 0)$.
  - The generalized family wise error rate: $gFWER(k) = Pr(V_n > k)$.
  - The per-comparison error rate: $PCER = E[V_n]/M$.
  - The per-family error rate: $PFER = E[V_n]$.
For a hypothesis testing, the $p$-value of a test is the smallest level $\alpha$ (constraint on the false alarm rate) at which $H_0$ would be rejected.

The smaller the $p$-value is, the more we want to decline $H_0$. 

$p$-Values
Unadjusted $p$-values

- Consider testing $M$ null hypotheses $\{H_0(m)\}$ ($m = 1, \ldots, M$) based on test statistics $\{T_n(m)\}$.

- The rejection region $C_n(m, \alpha)$ is chosen such that

  $$P_{Q_{0,m}}(T_n(m) \in C_n(m, \alpha)) \leq \alpha.$$ 

- The unadjusted $p$-value is defined as

  $$P_{0n}(m) = \inf\{\alpha \in [0, 1] : T_n(m) \in C_n(m, \alpha)\}.$$ 

- The null hypothesis $H_{0m}$ is rejected if $T_n(m) \in C_n(m, \alpha)$ or $P_{0n}(m) \leq \alpha$. 
Single Step and Stepwise MTP

- Single step MTP: each null hypothesis is tested using a rejection region that is independent of the results of the testing of other hypothesis.
- Stepwise MTP: the decision on a hypothesis is dependent on those of other hypotheses:
  - Step-down: The most significant null hypotheses are considered successively.
  - Step-up: The least significant null hypotheses are considered successively.
Family-wise Error Rate

- We require
  \[ FWER \leq \alpha. \]

- We use the p-values of individual tests to control the FWER. We assume
  \[ P(\hat{p}_i \leq u) \leq u. \]

- Bonferroni Procedure: If \( H_j \) is rejected when \( \hat{p}_j \leq \alpha/s \), for \( j = 1, \ldots, s \),
  then the FWER satisfies \( FWER \leq \alpha. \)

- In the Bonferroni procedure, the capability of detect cases in which \( H_j \) is false will be very low.
Holm Procedure

- In the Holm procedure, we sort the $p$-values. Suppose $\hat{p}_1 \leq ... \leq \hat{p}_s$. Then, we carry out the detection in the following steps:
  - Step 1. If $\hat{p}_1 \geq \alpha/s$, accept $H_1$, ..., $H_s$ and stop. If $\hat{p}_1 \leq \alpha/s$, reject $H_1$ and test the remaining $s - 1$ hypotheses at level $\alpha/(s - 1)$.
  - Step 2. If $\hat{p}_1 \leq \alpha/s$, but $\hat{p}_2 \geq \alpha/(s - 1)$, accept $H_2$, ..., $H_s$ and stop. If $\hat{p}_1 \leq \alpha/s$ and $\hat{p}_2 \leq \alpha/(s - 1)$, then, reject $H_1$ and $H_2$, and test the remaining hypotheses using level $\alpha/(s - 2)$.
  - ...

- Theorem: The Holm Procedure satisfies $FWER \leq \alpha$. 

Generic Stepdown Procedure

- Suppose that the test is based on a test statistic $T_{n,j}$ with large values indicating evidence against $H_j$. Suppose $T_{n,1} \geq \ldots \geq T_{n,s}$.

- **Step 1.** Let $K_1 = \{1, \ldots, s\}$. If $T_{n,1} \leq \hat{c}_{n,K_1}(1 - \alpha)$, then accept all hypotheses and stop; otherwise reject $H_1$ and continue.

- **Step 2.** Let $K_2 = \{2, \ldots, s\}$. If $T_{n,2} \leq \hat{c}_{n,K_2}(1 - \alpha)$, then accept all hypotheses and stop; otherwise reject $H_2$ and continue.

- ... 

- **Step s.** If $T_{n,s} \leq \hat{c}_{n,K_s}(1 - \alpha)$, then accept $H_s$; otherwise, reject $H_s$. 


Theorem: Let $P$ be the true distribution generating the data and $I(P)$ be the indices of the set of true hypotheses. Suppose that for any $I(P) \subseteq K$,

$$\hat{c}_{n,K}(1 - \alpha) \geq \hat{c}_{n,I(P)}(1 - \alpha),$$

then we have

$$FWER_P \leq P(\max(T_{n,j} : j \in I(P)) > \hat{c}_{n,I(P)}(1 - \alpha)).$$

Further, suppose that $\hat{c}_{n,K}(1 - \alpha)$ is used to test the intersection hypothesis $H_K$, then we have $FWER_P \leq \alpha$. 
Consider $N = 6033$ genes obtained from $n = 102$ men (50 normal control subjects and 52 prostate cancer patients). The goal is to find the genes whose expression levels differ between the two types of subjects.

$H_{j0}$ : gene $i$ is null; i.e., the expression level for gene $i$ has the same distribution for both groups.
Testing Statistic

- For gene $i$, we consider

$$t_i = \frac{\bar{x}_i(2) - \bar{x}_i(1)}{s_i},$$

where $s_i$ is the estimated standard error and $\bar{x}_i(j)$ is the average expression level in group $j$.

- We use the $z$-value of $t_i$:

$$z_i = \Phi^{-1}(F_{100}(t_i)),$$

where $\Phi$ and $F_{100}$ are the CDFs of standard normal and $t_{100}$ distributions. If we assume normal sampling and if $H_{0i}$ is true, we have

$$H_{0i} : z_i \sim \mathcal{N}(0, 1).$$
Existence of Non-null Ones

- The histogram of z-values is not normal, which implies the existence of some interesting non-null genes.
- If we use the Bonferroni bound, the rejection level is set to 0.05/6033, which is too conservative.
Bayesian Approach

- Two-groups model: We assume that the $N$ cases are each either null or non-null with prior probability $\pi_0$ or $\pi_1$. Usually we assume $\pi_0$ is large (say, $\pi_0 = 0.9$).

- We assume that $f_0(z)$ is standard Gaussian distributed, while $f_1(z)$ is some alternative density yielding $z$-values further away from 0. Then, we have a mixture density

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z).$$

- We apply the Bayesian rule and obtain the posterior probability upon receiving $z \in \mathcal{Z}$

$$\phi(\mathcal{Z}) = P\{null|\mathcal{Z}\} = \frac{\pi_0 F_0(\mathcal{Z})}{F_1(\mathcal{Z})},$$

where $F$ is the cumulative distribution function corresponding to $f$. 
Empirical Bayes Estimates

In practice, we do not know the prior mixture distribution $f$. In the situation of large data set, we can estimate $f$ (or $F$) empirically from the data:

$$
\bar{F}(Z) = \#\{z \in Z\} / N.
$$

"Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data." (proposed by H. E. Robbins)
In 1995, Benjamini and Hochberg proposed the concept of false discovery rate:

\[ Fdr_\mathcal{D} = \frac{\alpha_\mathcal{D}}{R_\mathcal{D}}, \]

where \( \mathcal{D} \) is the decision rule, \( R_\mathcal{D} \) is the number of cases rejected, and \( \alpha_\mathcal{D} \) is the number of cases that are falsely rejected.

The Fdr can be approximated by

\[ \overline{Fdr}(Z) = \frac{\pi_0 F_0(Z)}{\overline{F}(Z)}, \]

where \( Z \) is the region to reject the hypothesis.
FDR Control Algorithm

- We assume that the decision rule $D$ produces a $p$-value for each case $i$ such that $p_i$ has a uniform distribution.

- Benjamini and Hochberg proposed the following algorithm:
  - Sort the $p$-values of the hypotheses such that $p_1 \leq \ldots \leq p_N$.
  - Fix a value $q$ in $(0, 1)$. Find the index $i_{\text{max}}$ that is the largest index such that
    \[ p_i \leq \frac{i}{N} q. \]
  - Reject all $H_i$ such that $i \leq i_{\text{max}}$, and accept otherwise.
Empirical Bayesian Interpretation

- The B-H FDR control algorithm satisfies the following property: If the $p$-values corresponding to the correct null hypotheses are independent of each other, then the above rule controls the expected false discovery proportion at $q$, namely

\[
E \left[ Fdr_{BH}(q) \right] = \pi_0 q.
\]

- It can be interpreted in terms of empirical Bayesian approach:

\[
i_{\text{max}} = \max \{ i | Fdr(z_i) \leq q \}.
\]
Detection with Communication Constraints

- In many situations, the decision maker and the observers are not located at the same place. Communications are needed between the decision maker and the observers.


- We consider the simplest problem, namely the bivariate hypothesis testing when one of the variables is measured remotely and the information is transmitted over a noiseless channel with finite capacity.
Recap: Chernoff-Stein’s Lemma

- When the number of samples tends to infinity, we may have elegant asymptotic expressions for the performance of signal detection.

- Chernoff-Stein Lemma: Let $X_1, \ldots, X_n$ be i.i.d. and of distribution $Q$. Consider the hypothesis test between two alternatives, $Q = P_0$ and $Q = P_1$, where $D(P_0 || P_1) \leq \infty$ (the Kullback-Leibler distance). $A_n$ is the acceptance region of $H_1$. The error probabilities are

  $$P_F^n = P_0(A_n^c) \quad P_M^n = P_1(A_n).$$

Define $\beta(n, \epsilon) = \min_{A_n, P_F^n \leq \alpha} P_M^n$. Then, we have

$$\lim_{n \to \infty} \frac{1}{n} \log \beta(n, \epsilon) = -D(P_0 || P_1).$$
Problem Formulation

- We consider random variables $X$ and $Y$ taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$. $N$ samples are generated independently.

- We consider the hypothesis testing: $H_0$: $X$ and $Y$ are not mutually independent; $H_1$: $X$ and $Y$ are mutually independent.

- The samples for $Y$ can be obtained directly at the decision maker. But communications are needed to convey the samples of $X$ from the observer to the decision maker; i.e., instead of receiving $\{X_i\}$ directly, the decision maker receives $f(\{X_i\})$ where

$$\frac{1}{N} \log |f| \leq R,$$

where $|f|$ means the possible values of $f$. 
We define

\[ \beta_R(n, \epsilon) = \min_{\frac{1}{N} \log |f| \leq R} \beta(n, \epsilon, f), \]

where \( \beta(n, \epsilon, f) \) is the minimum missed detection rate if the coding scheme is \( f \).

We define

\[ \theta_k(R) = \sup_f \left\{ \frac{1}{k} D \left( P_{f(x_k)}, Y_k \parallel P_{f(x_k)} \times P_{Y_k} \right) \mid \log |f| \leq kR \right\}, \]

which is intuitive, and

\[ \theta(R) = \sup_k \theta_k(R). \]
Theorem: For every $R \geq 0$, we have
\[
\limsup_{n \to \infty} \frac{1}{n} \log \beta_R(n, \epsilon) \leq -\theta(R),
\]
for all $\epsilon \in (0, 1)$ and
\[
\lim_{\epsilon \to 0} \liminf_{n \to \infty} \frac{1}{n} \log \beta_R(n, \epsilon) \geq -\theta(R).
\]
The proof is based on the Chernoff-Stein’s Lemma.
Further Conclusions

- Can $\theta(R)$ be further simplified? The answer is positive.
- Does the following equality hold for every $R$ and $\epsilon \in (0, 1)$?

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_R(n, \epsilon) = -\theta(R).$$

The answer is also positive.
- What if $Y$ can also be compressed? Then, we have

$$\theta(R_X, R_Y) = \sup_{k, f, g} \left\{ \frac{1}{k} I(f(X^k); g(Y^k)) \mid \log |f| < kR_X, \log |f| < kR_Y \right\}.$$
More Generic Problem

- Consider two possible sequences of independent pairs \((X, Y)\) and \((\bar{X}, \bar{Y})\). We want to test

\[
H_0 : P_{X,Y} \quad \text{vs.} \quad H_1 : P_{\bar{X},\bar{Y}}.
\]

Again, we assume that \(Y\) can be observed directly while \(X\) needs to be sent from a remote observer.

- Similarly we can define

\[
\beta(n, \epsilon) = \min_{1 \leq |f| \leq R} \beta(n, \epsilon, f),
\]

where \(\beta(n, \epsilon, f)\) is the minimum missed detection rate if the coding scheme is \(f\), and

\[
\theta_k(R) = \sup_{f} \left\{ \frac{1}{k} D \left( P_{f(X^k), Y^k} \| P_{f(\bar{X}^k, \bar{Y}^k)} \right) \log |f| \leq kR \right\},
\]

which is intuitive, and \(\theta(R) = \sup_k \theta_k(R)\). The previous conclusions all hold.
Further Conclusions

- Can $\theta(R)$ be further simplified? The answer is positive. We can also obtain a lower bound:

$$\theta(R) \geq D(P_X||P_{\tilde{X}}) + D(P_Y||P_{\tilde{Y}}).$$

More advanced tools in information theory is needed.

- Does the following equality hold for every $R$ and $\epsilon \in (0, 1)$?

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_R(n, \epsilon) = -\theta(R).$$

The answer is also positive.
Tighter Bound


- Again, we consider the two hypotheses:
  
  \[ H_0 : P_{X,Y} \quad \text{vs.} \quad H_1 : P_{\bar{X},\bar{Y}}. \]

- We assume that the decision maker has the full information of \( Y \); \( X \) needs to be sent from a remote observer. Tighter bounds can be found than that of Ahlswede and Csiszar.
We consider the case in which the encoder of $X$ only outputs two possible messages (thus two bits), regardless of the data size of $X$. We denote this case by $R = 0_2$.

Theorem: For the case of complete data compression $R = 0_2$, for some $0 \leq \epsilon_0 \leq 1$ and all $0 < \epsilon < \epsilon_0$, we have

$$\theta(0_2, \epsilon) = \min_{\tilde{X}, \tilde{Y}} D(\tilde{X} \tilde{Y} || \bar{X} \bar{Y}),$$

where $\tilde{X}$ and $\tilde{Y}$ have the same marginal distributions as $X$ and $Y$.

The same conclusion holds when $Y$ also needs to be communicated, and $R_X = 0_2$ and $R_Y = 0_2$. 

Complete Data Compression

Again, they consider the case of zero rate transmission:

\[ R_X \to 0, \quad R_Y \to 0. \]
Main Conclusions

- For fixed-level simple hypothesis testing, it is shown that the two-sided one-bit compression/decision scheme is asymptotically optimal.

- In the case of composite hypothesis testing, the zero-rate compression may be (asymptotically suboptimal). The error exponent is not only a function of the null and alternative distribution classes, but also depends on the level $\epsilon$ and the sequences of the codebook sizes.