Asymptotic behaviour on Hypothesis Testing

zhaof17

May 2021

1 Hypothesis Testing

Consider the hypothesis testing problem,

$$\begin{cases} H = 0 & \text{(null)} \quad X \sim P \\ H = 1 & \text{(alternative)} \quad X \sim Q \end{cases} \tag{1}$$

Given one sample x, we will decide a test $\widehat{H} : \mathcal{X} \to \{0, 1\}$.

The type I error is defined as $\pi_0 = P(\widehat{H} = 1 | H = 0)$, also called false positive or false alarm.

The type II error is defined as $\pi_1 = P(\widehat{H} = 0 | H = 1)$, also called false negative or missing detection.

Let $A \triangleq \{x \in \mathcal{X} : \widehat{H}(x) = 1\}$ be the accept region for $H = 1$.

Theorem 1 (Neyman-Pearson Lemma)**.** *Let $A_\gamma = \{x : Q(x) > \gamma P(x)\}$ for $\gamma \in \mathbb{R}$. Then for any accept region A', if $\pi_1(A') < \pi_1(A_\gamma)$, then $\pi_0(A') > \pi_0(A_\gamma)$.*

A_γ is equivalent to log-likelihood ratio test.

For discrete $\mathcal{X}$, we have

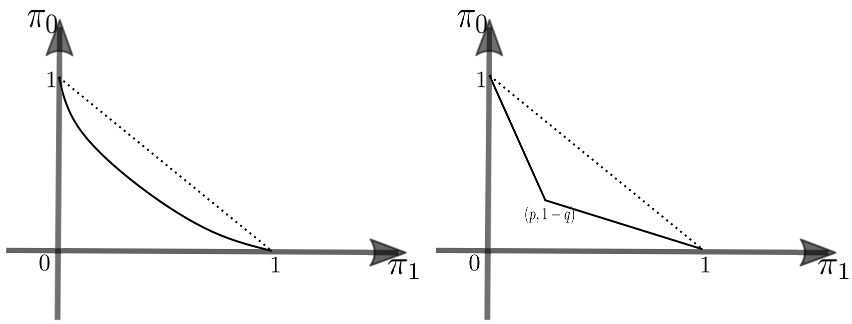
$$\pi_0 + \lambda \pi_1 \geq \lambda - \sum_{x \in \mathcal{X}} (P(x) - \lambda Q(x))^- \tag{2}$$

Theorem 1 implies that when π_1 decreases, π_0 increases.

If random decision rule is considered, for all decisions, the reachable $\pi_0 - \pi_1$ curve has the following shape in Fig. 1a The dotted line shows the random guess decision rule. That is for the operating point $(1 - p, p)$. The decision rule is simply determined by a random number generated from $\mathrm{Bern}(p)$ regardless of the data distribution P and Q. The solid curve can be reached by LRT test. Points in the regions between the two curves are reachable operating points, which perform better than random guess.

Proof of Theorem 1. We only need to show that

$$\pi_0(A') + \gamma \pi_1(A') \geq \pi_0(A_\gamma) + \gamma \pi_1(A_\gamma) \tag{3}$$



(a) $\pi_0 - \pi_1$ curve

(b) $\pi_0 - \pi_1$ curve for Example 1

Figure 1

This is because

$$
\begin{aligned}
\gamma\pi_0(A') + \pi_1(A') &= \gamma P(A') + Q(\bar{A}') \\
&= \gamma\,\mathbb{E}_P[\mathbb{1}_A(X)] + \mathbb{E}_Q[1 - \mathbb{1}_A(X)] \\
&= 1 + \gamma\,\mathbb{E}_P[\mathbb{1}_A(X)] - \mathbb{E}_P[\frac{Q(x)}{P(x)}\mathbb{1}_A(X)] \\
&= 1 + \mathbb{E}_P[\mathbb{1}_A(X)(\gamma - \frac{Q(x)}{P(x)})] \\
&\overset{(a)}{\geq} 1 + \mathbb{E}_P[\mathbb{1}_{A'}(X)(\gamma - \frac{Q(x)}{P(x)})] \\
&= \pi_0(A_\gamma) + \gamma\pi_1(A_\gamma)
\end{aligned}
$$

where (a) holds since the set A' is the maximum set which minimizes $\sum_{x\in A} P(x)(\gamma - \frac{Q(x)}{P(x)})$. $\square$

Example 1. *Consider hypothesis testing between two Bernoulli random variables,* $P = \mathrm{Bern}(p), Q = \mathrm{Bern}(q)$ *such that* $p < \frac{1}{2} < q$*. We can verify that the solid curve in Fig. 1a becomes a piece-wise linear line with turning point* $(p, 1-q)$*.*

2 Multiple samples

In this section we consider the two types of error when multiple samples are observed. Let $x^n = (x_1, \ldots, x_n)$ be i.i.d. sampled from either P or Q.

$$
\begin{cases}
H = 0 \quad \text{(null)} \quad X^n \sim P^n \\
H = 1 \quad \text{(alternative)} \quad X^n \sim Q^n
\end{cases}
\tag{4}
$$

where P^n or Q^n represents the joint distribution of distribution of X^n. Then the LRT is rewritten as $\frac{1}{n}\sum_{i=1}^n \ell(x_i)$ where $\ell(x) = \log\frac{Q(x)}{P(x)}$. The decision rule for LRT is $A_\gamma \triangleq \{x^n : \frac{1}{n}\sum_{i=1}^n \ell(x_i) > \gamma\}$, and the two types of error are given by

$$\pi_0^{(n)} = P^n(A_\gamma)$$
$$\pi_1^{(n)} = Q^n(\bar{A}_\gamma)$$

For given γ, we first analyze the decaying rate of $\pi_0^{(n)}$ and $\pi_1^{(n)}$. We define the decay exponent as

$$E_i \triangleq -\lim_{n\to\infty} \frac{1}{n} \log \pi_i^{(n)} \text{ for } i = 0, 1 \tag{5}$$

Then $\pi_0^{(n)} \doteq \exp(-nE_0)$ and $\pi_1^{(n)} \doteq \exp(-nE_1)$.

2.1 Analysis of exponent with Cramér's Theorem

The analysis is based on the Cramér's Theorem and is applied to the sample mean of $\ell(x_i)$. To make both two types of error decay, we should require γ lies in between $(-D(P||Q), D(Q||P))$. The end point be obtained by computing $\mathbb{E}_P[l(X)]$ and $\mathbb{E}_Q[l(X)]$ directly. $\gamma = -D(P||Q)$ implies $\pi_0^{(n)} \to 1$ while $\gamma = D(Q||P)$ implies $\pi_1^{(n)} \to 1$.

For the non-trivial case, that is, $-D(P||Q) < \gamma < D(Q||P)$, we can quickly obtain that $E_0 = \psi_P^*(\gamma)$ and $E_1 = \psi_Q^*(\gamma)$.

Let the log-MGF $\psi_P(\lambda) \triangleq \log \mathbb{E}_P[\exp(\lambda\ell(X))] = \log \sum_{x\in\mathcal{X}}[P(X)]^{1-\lambda}[Q(x)]^\lambda$. Using $Q(x) = P(x)e^{\ell(x)}$, we can obtain $\psi_Q(\lambda) \triangleq \log \mathbb{E}_Q[\exp(\lambda\ell(X))] = \log \mathbb{E}_P[\exp((\lambda+1)\ell(X))] = \psi_P(\lambda+1)$. Then $\psi_P^*(\gamma) \triangleq \sup_{\lambda\in\mathbb{R}}[\lambda\gamma - \psi_P(\lambda)]$. The optimal λ is chosen to satisfy $\psi_P'(\lambda) = \gamma$. The relation can be rewritten as $\psi_P'(\lambda) = \frac{\mathbb{E}_P[\ell(X)\exp(\lambda\ell(X))]}{\mathbb{E}_P[\exp(\lambda\ell(X))]} = \mathbb{E}_P[\ell(X)\exp(\lambda\ell(X) - \psi_P(\lambda))] = \mathbb{E}_{P^{(\lambda)}}[\ell(X)]$ where the geometric mixture distribution $P^{(\lambda)}$ is defined as

$$P^{(\lambda)}(x) = P(x)\exp(\lambda\ell(X) - \psi_P(\lambda)) = \frac{[P(X)]^{1-\lambda}[Q(x)]^\lambda}{\sum_{x\in\mathcal{X}}[P(X)]^{1-\lambda}[Q(x)]^\lambda} \tag{6}$$

When $\lambda = 0$, $P^{(\lambda)} = P$ and $\gamma = -D(P||Q)$; When $\lambda = 1$, $P^{(\lambda)} = Q$ and $\gamma = D(Q||P)$. Besides, γ is an monotonically increasing function of λ from the property of the conjugate log-MGF. Therefore, the domain of definition for λ is $(0, 1)$ to guarantee $-D(P||Q) < \gamma < D(Q||P)$.

$\psi_Q^*(\gamma)$ can be expressed in term of $\psi_P^*(\gamma)$: $\psi_Q^*(\gamma) = \sup_{\lambda\in\mathbb{R}}[\lambda\gamma - \psi_P(\lambda+1)] = \psi_P^*(\gamma) - \gamma$.

We can draw the function $\psi_P(\lambda)$ and illustrate it in Fig. 2. At the point $(\lambda_0, \psi_P(\lambda)$, the slope of the tangent line is γ such that $\psi_P'(\lambda_0) = \gamma$. Since $\psi_P^*(\gamma) = \gamma\lambda_0 - \psi_P(\lambda_0)$. The geometric meaning of E_0 is the length of the intercept for the tangent line, and E_1 is the length of y-axis of the intersection between the tangent line and $\lambda = 1$. Fig. 2 also shows a right trapezoid whose two bases have length E_0 and E_1. the lateral side has length 1 on the right angle side. Finally, Fig. 2 shows the trade-off between E_0 and E_1. One error increases while the other decreases.

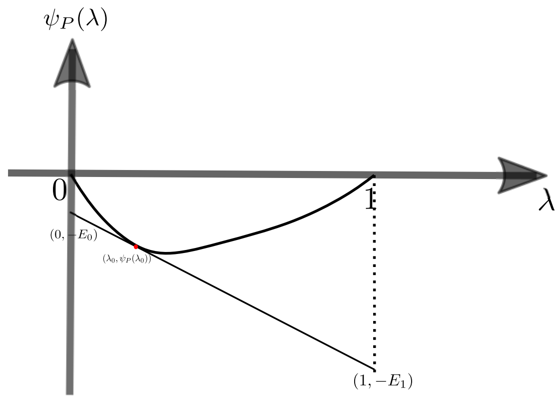


Figure 2: log-MGF of $\ell(X)$

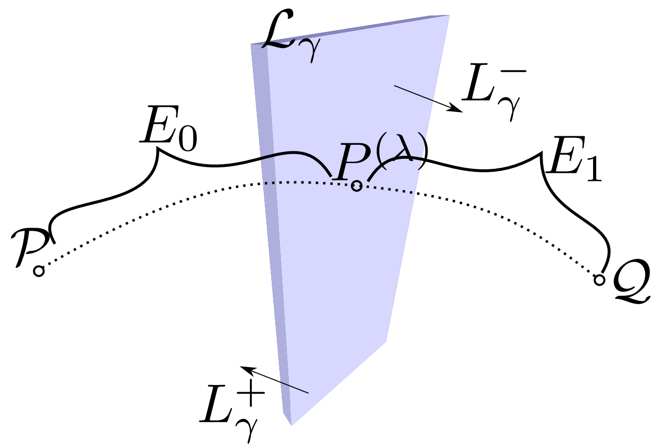


Figure 3: E_0, E_1 representation in distribution space

2.2 Analysis of exponent with Sanov's Theorem

Since we consider discrete random variables, Sanov's theorem can be applied to obtain another illustration for E_0 and E_1. That is $E_0 = \min_{R \in \mathcal{L}_\gamma} D(R||P)$ and $E_1 = \min_{R \in \mathcal{L}_\gamma} D(R||Q)$ where R is the empirical distribution on the hyper-plane $\mathcal{L}_\gamma \triangleq \{R : \mathbb{E}_R[\ell(X)] = \gamma\}$. We use L_γ^+ to represent the space when $\mathbb{E}_R[\ell(X)] > \gamma$. L_γ^- is defined similarly. By Lagrange's method, we can find a distribution $P^{(\lambda)}$ defined in (6) and belongs to $\mathcal{L}_\gamma$. As we change γ from $-D(P||Q)$ to $D(Q||P)$, λ changes from 0 to 1, and the distribution $P^{(\lambda)}$ changes from P to Q. the change path for $P^{(\lambda)}$ can be draw in distribution space as shown in Fig. 3. Then the two error exponents are the distances $E_0 = D(P^{(\lambda)}||P)$ and $E_1 = D(P^{(\lambda)}||Q)$.

There are two ways to study the optimal decay rate of $\pi_0^{(n)}$ and $\pi_1^{(n)}$. One way is calculating the fast exponential rate of one error while controlling the other error within a given threshold. Or we try to minimize the weighted sum of the two types of error under Bayesian setting. The first way is quantified by Chernoff-Stein Lemma, while the second is studied by Chernoff information.

Theorem 2 (Chernoff-Stein Lemma)**.** *Let $\pi_0^{(n)} < \epsilon$, and the optimal type II error $\hat{\pi}_1^{(n)}$ is defined as*

$$\hat{\pi}_1^{(n)} \triangleq \min_{A:\pi_0^{(n)}(A)<\epsilon} \pi_1^{(n)}(A)$$

. Then $\lim_{n\to\infty} \frac{1}{n} \log \hat{\pi}_1^{(n)} = -D(P||Q)$.

From N-P Lemma, we should choose LLR test to achieve the minimal $\hat{\pi}_1^{(n)}$. If $E_0 > 0$, when n is sufficiently large, we have $\pi_0^{(n)} < \epsilon$. Then from the analysis of the last section, when we let $E_0 > 0$ to be arbitrarily small, we can achieve E_1 arbitrarily close to $D(P||Q)$. That is, we have $D(P||Q) \leq E_1 < D(P||Q) - \delta$ for any $\delta > 0$.

Theorem 3 (Chernoff Information)**.** *Consider the error probability defined as $P_e^{(n)} = P(\widehat{H} \neq H) = P(\widehat{H} = 1)\pi_0^{(n)} + P(\widehat{H} = 0)\pi_1^{(n)}$. Then*

$$\inf \liminf_{n\to\infty} \frac{1}{n} \log P_e^{(n)} = -\psi_P^*(0) \tag{7}$$

where $\psi_P^*(0) = -\min_{\lambda\in[0,1]} \log \sum_{x\in\mathcal{X}} [P(X)]^{1-\lambda}[Q(x)]^\lambda$

Notice that $\psi_P^*(0)$ corresponds to $\gamma = 0$, which means the slope equals to zero in Fig. 3. This limit is achieved for LRT with λ satisfying $\psi'_P(\lambda) = 0$. In such case, $E_0 = E_1 = \psi_P^*(0)$. For the error of other detection method, $P_e^{(n)} \doteq \exp(-n \min\{E_0, E_1\})$. By NP Lemma, when adopting decision rules other than LRT, $\min\{E_0, E_1\}$ will become smaller. Thus left hand side in (7) becomes larger. As a result, we only need to consider LRT, that is, to maximize $\min\{E_0, E_1\}$. From Fig. 3, the maximization is achieved when $\gamma = \psi'_P(\lambda) = 0$.