Chapter 8

Hypothesis Testing

Po-Ning Chen

Department of Communications Engineering

National Chiao-Tung University

Hsin Chu, Taiwan 30050

Error exponent and divergence

Definition 8.1 (exponent) A real number a is said to be the *exponent* for a sequence of non-negative quantities $\{a_n\}_{n\geq 1}$ converging to zero, if

$$a = \lim_{n\to\infty} \left(-\frac{1}{n}\log a_n\right).$$

- In operation, exponent is an index for the exponential rate-of-convergence for sequence a_n. We can say that for any $\gamma > 0$,

  $$e^{-n(a+\gamma)} \leq a_n \leq e^{-n(a-\gamma)},$$

  as n large enough.

- Recall that in proving the channel coding theorem, the probability of decoding error for channel block codes can be made arbitrarily close to zero when the rate of the codes is less than channel capacity.

- Actually, this result can be mathematically written as:

  $$P_e(\mathscr{C}_n^*) \to 0 \text{ as } n \to \infty,$$

  provided $R = \limsup_{n\to\infty}(1/n)\log\|\mathscr{C}_n^*\| < C$, where $\mathscr{C}_n^*$ is the optimal code for block length n.

- From the theorem, we only know the decoding error will vanish as block length increases; but, it does not reveal how fast the decoding error approaches zero.

Error exponent and divergence

- In other words, we do not know the rate-of-convergence of the decoding error. Sometimes, this information is very important, especially for one to decide the sufficient block length to achieve some error bound.
- The first step of investigating the rate-of-convergence of the decoding error is to compute its *exponent*, if the decoding error decays to zero exponentially fast (it indeed does for memoryless channels.) This exponent, as a function of the rate, is in fact called the *channel reliability function*, and will be discussed in the next chapter.
- For the hypothesis testing problems, the type II error probability of fixed test level also decays to zero as the number of observations increases. As it turns out, its *exponent* is the *divergence* of the null hypothesis distribution against alternative hypothesis distribution.

Lemma 8.2 (Stein's lemma) For a sequence of i.i.d. observations X^n which is possibly drawn from either null hypothesis distribution P_{X^n} or alternative hypothesis distribution $P_{\hat{X}^n}$, the type II error satisfies

$$(\forall\ \varepsilon \in (0,1)) \quad \lim_{n\to\infty} -\frac{1}{n}\log \beta_n^*(\varepsilon) = D(P_X\|P_{\hat{X}}),$$

where $\beta_n^*(\varepsilon) = \min_{\alpha_n \le \varepsilon} \beta_n$, and α_n and β_n represent the type I and type II errors respectively.

Proof: [1. *Forward Part*]

In the forward part, we prove that there exists an acceptance region for null hypothesis such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log \beta_n(\varepsilon) \ge D(P_X\|P_{\hat{X}}).$$

step 1: divergence typical set. For any $\delta > 0$, define divergence typical set as

$$\mathcal{A}_n(\delta) \stackrel{\triangle}{=} \left\{ x^n \ : \ \left| \frac{1}{n}\log \frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} - D(P_X\|P_{\hat{X}}) \right| < \delta \right\}.$$

Note that in divergence typical set,

$$P_{\hat{X}^n}(x^n) \le P_{X^n}(x^n) e^{-n(D(P_X\|P_{\hat{X}})-\delta)}.$$

step 2: computation of type I error. By weak law of large number,

$$P_{X^n}(\mathcal{A}_n(\delta)) \to 1.$$

Hence,

$$\alpha_n = P_{X^n}(\mathcal{A}_n^c(\delta)) < \varepsilon,$$

for sufficiently large n.

step 3: computation of type II error.

$$\begin{aligned}
\beta_n(\varepsilon) &= P_{\hat{X}^n}(\mathcal{A}_n(\delta)) \\
&= \sum_{x^n \in \mathcal{A}_n(\delta)} P_{\hat{X}^n}(x^n) \\
&\leq \sum_{x^n \in \mathcal{A}_n(\delta)} P_{X^n}(x^n) e^{-n(D(P_X\|P_{\hat{X}})-\delta)} \\
&\leq e^{-n(D(P_X\|P_{\hat{X}})-\delta)} \sum_{x^n \in \mathcal{A}_n(\delta)} P_{X^n}(x^n) \\
&\leq e^{-n(D(P_X\|P_{\hat{X}})-\delta)}(1-\alpha_n).
\end{aligned}$$

Hence,

$$-\frac{1}{n}\log\beta_n(\varepsilon) \geq D(P_X\|P_{\hat{X}}) - \delta + \frac{1}{n}\log(1-\alpha_n),$$

which implies

$$\liminf_{n\to\infty} -\frac{1}{n}\log \beta_n(\varepsilon) \geq D(P_X\|P_{\hat{X}}) - \delta.$$

The above inequality is true for any $\delta > 0$. Therefore

$$\liminf_{n\to\infty} -\frac{1}{n}\log_2 \beta_n(\varepsilon) \geq D(P_X\|P_{\hat{X}}).$$

[2. *Converse Part*]

In the converse part, we will prove that for any acceptance region for null hypothesis $\mathcal{B}_n$ satisfying the type I error constraint, i.e.,

$$\alpha_n(\mathcal{B}_n) = P_{X^n}(\mathcal{B}_n^c) \leq \varepsilon,$$

its type II error $\beta_n(\mathcal{B}_n)$ satisfies

$$\limsup_{n\to\infty} -\frac{1}{n}\log \beta_n(\mathcal{B}_n) \leq D(P_X\|P_{\hat{X}}).$$

$$\begin{aligned}
\beta_n(\mathcal{B}_n) = P_{\hat{X}^n}(\mathcal{B}_n) &\geq P_{\hat{X}^n}(\mathcal{B}_n \cap \mathcal{A}_n(\delta)) \\
&\geq \sum_{x^n \in \mathcal{B}_n \cap \mathcal{A}_n(\delta)} P_{\hat{X}^n}(x^n) \\
&\geq \sum_{x^n \in \mathcal{B}_n \cap \mathcal{A}_n(\delta)} P_{X^n}(x^n) e^{-n(D(P_X \| P_{\hat{X}}) + \delta)} \\
&= e^{-n(D(P_X \| P_{\hat{X}}) + \delta)} P_{X^n}(\mathcal{B}_n \cap \mathcal{A}_n(\delta)) \\
&\geq e^{-n(D(P_X \| P_{\hat{X}}) + \delta)} \left(1 - P_{X^n}(\mathcal{B}_n^c) - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right) \\
&\geq e^{-n(D(P_X \| P_{\hat{X}}) + \delta)} \left(1 - \alpha_n(\mathcal{B}_n) - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right) \\
&\geq e^{-n(D(P_X \| P_{\hat{X}}) + \delta)} \left(1 - \varepsilon - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right).
\end{aligned}$$

Hence,

$$-\frac{1}{n} \log \beta_n(\mathcal{B}_n) \leq D(P_X \| P_{\hat{X}}) + \delta + \frac{1}{n} \log \left(1 - \varepsilon - P_{X^n}\left(\mathcal{A}_n^c(\delta)\right)\right),$$

which implies that

$$\limsup_{n \to \infty} -\frac{1}{n} \log \beta_n(\mathcal{B}_n) \leq D(P_X \| P_{\hat{X}}) + \delta.$$

The above inequality is true for any $\delta > 0$. Therefore,

$$\limsup_{n \to \infty} -\frac{1}{n} \log \beta_n(\mathcal{B}_n) \leq D(P_X \| P_{\hat{X}}).$$

Composition of sequence of i.i.d. observations

- Stein's lemma gives the exponent of the type II error probability for fixed test level.
- As a result, this exponent, which is the divergence of null hypothesis distribution against alternative hypothesis distribution, is independent of the type I error bound ε for i.i.d. observations.
- Specifically under i.i.d. environment, the probability for each sequence of x^n depends only on its *composition*, which is defined as an $|\mathcal{X}|$-dimensional vector, and is of the form

$$\left(\frac{\#1(x^n)}{n}, \frac{\#2(x^n)}{n}, \ldots, \frac{\#k(x^n)}{n}\right),$$

  where $\mathcal{X} = \{1, 2, \ldots, k\}$, and $\#i(x^n)$ is the number of occurrences of symbol i in x^n.
- The probability of x^n is therefore can be written as

$$P_{X^n}(x^n) = P_X(1)^{\#1(x^n)} \times P_X(2)^{\#2(x^n)} \times P_X(k)^{\#k(x^n)}.$$

- Note that $\#1(x^n) + \cdots + \#k(x^n) = n$.
- Since the composition of sequence decides its probability deterministically, all sequences with the same composition should have the same statistical property, and hence should be treated the same when processing.

Composition of sequence of i.i.d. observations

- Instead of manipulating the sequences of observations based on the *typical-set*-like concept, we may focus on their *compositions*.
- As it turns out, such approach yields simpler proofs and better geometrical explanations for theories under i.i.d. environment.
- (It needs to be pointed out that for cases when *composition* alone can not decide the probability, this viewpoint does not seem to be effective.)

Lemma 8.3 (polynomial bound on number of composition) The number of compositions increases polynomially fast, while the number of possible sequences increases exponentially fast.

Proof:

- Let $\mathcal{P}_n$ denotes the set of all possible compositions.
- $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$

□

Lemma 8.4 (probability of sequences of the same composition) The probability of the sequences of composition $\mathcal{C}$ with respect to distribution P_{X^n} satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}|}} e^{-nD(P_{\mathcal{C}}\|P_X)} \leq P_{X^n}(\mathcal{C}) \leq e^{-nD(P_{\mathcal{C}}\|P_X)},$$

where $P_{\mathcal{C}}$ is the composition distribution for composition $\mathcal{C}$, and $\mathcal{C}$ (by abusing notation without ambiguity) is also used to represent the set of all sequences (in $\mathcal{X}^n$) of composition $\mathcal{C}$.

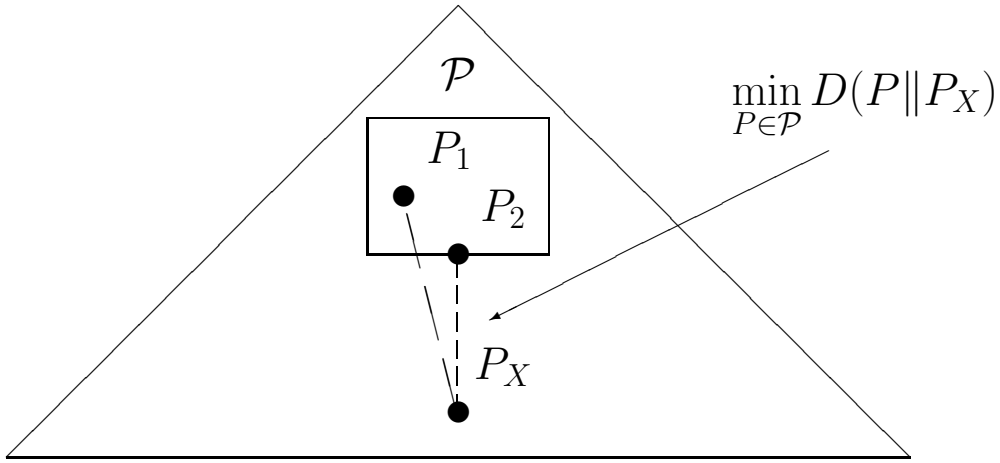
Theorem 8.5 (Sanov's Theorem) Let $\mathcal{E}_n$ be the set that consists of all compositions over finite alphabet $\mathcal{X}$, whose composition distribution belongs to $\mathcal{P}$. Fix a sequence of product distribution $P_{X^n} = \prod_{i=1}^n P_X$. Then,

$$\liminf_{n\to\infty} -\frac{1}{n} \log P_{X^n}(\mathcal{E}_n) \geq \inf_{P_{\mathcal{C}}\in\mathcal{P}} D(P\|P_X),$$

where $P_{\mathcal{C}}$ is the composition distribution for composition $\mathcal{C}$. If, in addition, for every distribution P in $\mathcal{P}$, there exists a sequence of composition distributions $P_{\mathcal{C}_1}, P_{\mathcal{C}_2}, P_{\mathcal{C}_3}, \ldots \in \mathcal{P}$ such that $\limsup_{n\to\infty} D(P_{\mathcal{C}_n}\|P_X) = D(P\|P_X)$, then

$$\limsup_{n\to\infty} -\frac{1}{n} \log P_{X^n}(\mathcal{E}_n) \leq \inf_{P\in\mathcal{P}} D(P_{\mathcal{C}}\|P_X).$$

Geometrical interpretation for Sanov's theorem



The geometric meaning for Sanov's theorem.

Example 8.6 • **Question:** One wants to roughly estimate the probability that the average of the throws is greater or equal to 4, when tossing a fair dice n times.

- He observes that whether the requirement is satisfied only depends on the compositions of the observations.
- Let $\mathcal{E}_n$ be the set of compositions which satisfy the requirement.

$$\mathcal{E}_n = \left\{ \mathcal{C} \ : \ \sum_{i=1}^{6} i P_{\mathcal{C}}(i) \geq 4 \right\}.$$

- To minimize $D(P_{\mathcal{C}} \| P_X)$ for $\mathcal{C} \in \mathcal{E}_n$, we can use the Lagrange multiplier technique (since divergence is convex with respect to the first argument.) with the constraints on $P_{\mathcal{C}}$ being:

$$\sum_{i=1}^{6} i P_{\mathcal{C}}(i) = k \quad \text{and} \quad \sum_{i=1}^{6} P_{\mathcal{C}}(i) = 1$$

for $k = 4, 5, 6, \ldots, n$.

- So it becomes to minimize:

$$\sum_{i=1}^{6} P_{\mathcal{C}}(i) \log \frac{P_{\mathcal{C}}(i)}{P_X(i)} + \lambda_1 \left(\sum_{i=1}^{6} i P_{\mathcal{C}}(i) - k \right) + \lambda_2 \left(\sum_{i=1}^{6} P_{\mathcal{C}}(i) - 1 \right).$$

Geometrical interpretation for Sanov's theorem

- By taking the derivatives, we found that the minimizer should be of the form

$$P_{\mathcal{C}}(i) = \frac{e^{\lambda_1 \cdot i}}{\sum_{j=1}^{6} e^{\lambda_1 \cdot j}},$$

  for λ_1 is chosen to satisfy

$$\sum_{i=1}^{6} i P_{\mathcal{C}}(i) = k. \tag{8.1.1}$$

- Since the above is true for all $k \geq 4$, it suffices to take the smallest one as our solution, i.e., $k = 4$.

- Finally, by solving (8.1.1) for $k = 4$ numerically, the minimizer should be

$$P_{\mathcal{C}^*} = (0.1031, 0.1227, 0.1461, 0.1740, 0.2072, 0.2468),$$

  and the exponent of the desired probability is $D(P_{\mathcal{C}^*} \| P_X) = 0.0433$ nat.

- Consequently,

$$P_{X^n}(\mathcal{E}_n) \approx e^{-0.0433 \cdot n}.$$

Divergence typical set on composition

- Divergence typical set in Stein's lemma:

$$\mathcal{A}_n(\delta) \stackrel{\triangle}{=} \left\{ x^n \ : \ \left| \frac{1}{n} \log \frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} - D(P_X \| P_{\hat{X}}) \right| < \delta \right\}.$$

- $$\mathcal{T}_n(\delta) \stackrel{\triangle}{=} \{ x^n \in \mathcal{X}^n \ : \ D(P_{\mathcal{C}_{x^n}} \| P_X) \leq \delta \},$$

  where $\mathcal{C}_{x^n}$ represents the composition of x^n.

- $P_{X^n}(\mathcal{T}_n(\delta)) \to 1$ is justified by

$$\begin{aligned}
1 - P_{X^n}(\mathcal{T}_n(\delta)) &= \sum_{\{C \ : \ D(P_{\mathcal{C}} \| P_X) > \delta\}} P_{X^n}(\mathcal{C}) \\
&\leq \sum_{\{C \ : \ D(P_{\mathcal{C}} \| P_X) > \delta\}} e^{-nD(P_{\mathcal{C}} \| P_X)}, \text{ from Lemma 8.4.} \\
&\leq \sum_{\{C \ : \ D(P_{\mathcal{C}} \| P_X) > \delta\}} e^{-n\delta} \\
&\leq (n+1)^{|\mathcal{X}|} e^{-n\delta}, \text{ cf. Lemma 8.3.}
\end{aligned}$$

Universal source coding on composition

- Universal code

$$f_n \ : \ \mathcal{X}^n \to \bigcup_{i=1}^{\infty} \{0,1\}^i$$

for i.i.d. source:

$$\frac{1}{n} \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \ell(f_n(x^n)) \to H(X),$$

as n goes to infinity.

Example 8.7 (universal encoding using compositions)

- Binary-index the compositions using $\log_2(n+1)^{|\mathcal{X}|}$ bits, and denote this binary index for composition $\mathcal{C}$ by $a(\mathcal{C})$.
  - Let $\mathcal{C}_{x^n}$ denote the composition with respect to x^n, i.e. $x^n \in \mathcal{C}_{x^n}$.
- Binary-index the elements in $\mathcal{C}$ using $n \cdot H(P_{\mathcal{C}})$ bits, and denote this binary index for elements in $\mathcal{C}$ by $b(\mathcal{C}_{x^n})$.
  - For each composition $\mathcal{C}$, we know that the number of sequence x^n in $\mathcal{C}$ is at most $2^{n \cdot H(P_{\mathcal{C}})}$ (Here, $H(P_{\mathcal{C}})$ is measured in bits. I.e., the logarithmic base in entropy is 2. See the proof of Lemma 8.4).

Universal source coding on composition

- Define a universal encoding function f_n as

$$f_n(x^n) = \text{concatenation}\{a(C_{x^n}), b(C_{x^n})\}.$$

- Then this encoding rule is a universal code for all i.i.d. sources.

Proof:

$$\begin{aligned}
\bar{\ell}_n &= \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)\ell(a(\mathcal{C}_{x^n})) + \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n)\ell(b(\mathcal{C}_{x^n})) \\
&\leq \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \cdot \log_2(n+1)^{|\mathcal{X}|} + \sum_{x^n \in \mathcal{X}^n} P_{X^n}(x^n) \cdot n \cdot H(P_{\mathcal{C}_{x^n}}) \\
&\leq |\mathcal{X}| \cdot \log_2(n+1) + \sum_{\{\mathcal{C}\}} P_{X^n}(\mathcal{C}) \cdot n \cdot H(P_{\mathcal{C}}).
\end{aligned}$$

$$\frac{1}{n}\bar{\ell}_n \leq \frac{|\mathcal{X}| \times \log_2(n+1)}{n} + \sum_{\{\mathcal{C}\}} P_{X^n}(\mathcal{C}) H(P_{\mathcal{C}}).$$

$$
\begin{aligned}
& \sum_{\{\mathcal{C}\}} P_{X^n}(\mathcal{C}) H(P_{\mathcal{C}}) \\
= & \sum_{\{\mathcal{C}\in\mathcal{T}_n(\delta)\}} P_{X^n}(\mathcal{C}) H(P_{\mathcal{C}}) + \sum_{\{\mathcal{C}\notin\mathcal{T}_n(\delta)\}} P_{X^n}(\mathcal{C}) H(P_{\mathcal{C}}) \\
\leq & \max_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)\leq\delta/\log(2)\}} H(P_{\mathcal{C}}) + \sum_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)>\delta/\log(2)\}} P_{X^n}(\mathcal{C}) H(P_{\mathcal{C}}) \\
\leq & \max_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)\leq\delta/\log(2)\}} H(P_{\mathcal{C}}) + \sum_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)>\delta/\log(2)\}} 2^{-nD(P_{\mathcal{C}}\|P_X)} H(P_{\mathcal{C}}), \\
& \text{(From Lemma 8.4)} \\
\leq & \max_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)\leq\delta/\log(2)\}} H(P_{\mathcal{C}}) + \sum_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)>\delta/\log(2)\}} e^{-n\delta} H(P_{\mathcal{C}}) \\
\leq & \max_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)\leq\delta/\log(2)\}} H(P_{\mathcal{C}}) + \sum_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)>\delta/\log(2)\}} e^{-n\delta} \log_2|\mathcal{X}| \\
\leq & \max_{\{\mathcal{C}\,:\,D(P_{\mathcal{C}}\|P_X)\leq\delta/\log(2)\}} H(P_{\mathcal{C}}) + (n+1)^{|\mathcal{X}|} e^{-n\delta} \log_2|\mathcal{X}|,
\end{aligned}
$$

where the second term of the last step vanishes as $n \to \infty$. (Note that when base-2 logarithm is taken in divergence instead of natural logarithm, the range $[0, \delta]$ in

Universal source coding on composition

$\mathcal{T}_n(\delta)$ should be replaced by $[0, \delta/\log(2)]$.) It remains to show that

$$\max_{\{\mathcal{C} \,:\, D(P_{\mathcal{C}} \| P_X) \leq \delta/\log(2)\}} H(P_{\mathcal{C}}) \leq H(X) + \gamma(\delta),$$

where $\gamma(\delta)$ only depends on δ, and approaches zero as $\delta \longrightarrow 0$.

$\ldots$ □

Likelihood ratio versus divergence

- Recall that the Neyman-Pearson lemma indicates that the optimal test for two hypothesis is of the form

$$\frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} \underset{<}{\overset{>}{}} \tau. \tag{8.1.2}$$

- This is the likelihood ratio test and the quantity $P_{X^n}(x^n)/P_{\hat{X}^n}(x^n)$ is called the *likelihood ratio*.

- If a log operation is performed on both sides of (8.1.2), the test remains.

Likelihood ratio versus divergence

$$\begin{aligned}
\log \frac{P_{X^n}(x^n)}{P_{\hat{X}^n}(x^n)} &= \sum_{i=1}^{n} \log \frac{P_X(x_i)}{P_{\hat{X}}(x_i)} \\
&= \sum_{a \in \mathcal{X}} [\#a(x^n)] \log \frac{P_X(a)}{P_{\hat{X}}(a)} \\
&= \sum_{a \in \mathcal{X}} [nP_{\mathcal{C}_{x^n}}(a)] \log \frac{P_X(a)}{P_{\hat{X}}(a)} \\
&= n \cdot \sum_{a \in \mathcal{X}} P_{\mathcal{C}_{x^n}}(a) \log \frac{P_X(a)}{P_{\mathcal{C}_{x^n}}(a)} \frac{P_{\mathcal{C}_{x^n}}(a)}{P_{\hat{X}}(a)} \\
&= n \left[\sum_{a \in \mathcal{X}} P_{\mathcal{C}_{x^n}}(a) \log \frac{P_{\mathcal{C}_{x^n}}(a)}{P_{\hat{X}}(a)} - \sum_{a \in \mathcal{X}} P_{\mathcal{C}_{x^n}}(a) \log \frac{P_{\mathcal{C}_{x^n}}(a)}{P_X(a)} \right] \\
&= n \left[D(P_{\mathcal{C}_{x^n}} \| P_{\hat{X}}) - D(P_{\mathcal{C}_{x^n}} \| P_X) \right]
\end{aligned}$$

Hence, (8.1.2) is equivalent to

$$D(P_{\mathcal{C}_{x^n}} \| P_{\hat{X}}) - D(P_{\mathcal{C}_{x^n}} \| P_X) \gtrless \frac{1}{n} \log \tau. \tag{8.1.3}$$

- This equivalence means that for hypothesis testing, selection of the acceptance region can be made upon *compositions* instead of *observations*.

Likelihood ratio versus divergence

- In other words, the optimal decision function can be defined as:

$$\phi(\mathcal{C}) = \begin{cases} 0, & \text{if composition } \mathcal{C} \text{ is classified to belong to null hypothesis} \\ & \text{according to (8.1.3);} \\ 1, & \text{otherwise.} \end{cases}$$

Exponent of Bayesian cost

- Randomization is of no help to Bayesian test.

$$\phi(x^n) \;=\; \begin{cases} 0, & \text{with probability } \eta; \\ 1, & \text{with probability } 1-\eta; \end{cases}$$

  satisfies

$$\pi_0 \eta P_{X^n}(x^n) + \pi_1 (1-\eta) P_{\hat{X}^n}(x^n) \geq \min\{\pi_0 P_{X^n}(x^n), \pi_1 P_{\hat{X}^n}(x^n)\}.$$

- Now suppose the acceptance region for null hypothesis is

$$\mathcal{A} \stackrel{\triangle}{=} \{\mathcal{C} \;:\; D(P_{\mathcal{C}} \| P_{\hat{X}}) - D(P_{\mathcal{C}} \| P_X) > \tau'\}.$$

- Then by Sanov's theorem, the exponent of type II error, β_n, is

$$\min_{\mathcal{C} \in \mathcal{A}} D(P_{\mathcal{C}} \| P_{\hat{X}}).$$

- Similarly, the exponent of type I error, α_n is

$$\min_{\mathcal{C} \in \mathcal{A}^c} D(P_{\mathcal{C}} \| P_X).$$

Exponent of Bayesian cost

- Lagrange multiplier: by taking derivative of

$$D(P_{\widetilde{X}}\|P_{\hat{X}}) + \lambda(D(P_{\widetilde{X}}\|P_{\hat{X}}) - D(P_{\widetilde{X}}\|P_X) - \tau') + \nu\left(\sum_{x\in\mathcal{X}} P_{\widetilde{X}}(x) - 1\right)$$

  with respective to each $P_{\widetilde{X}}(x)$, we have

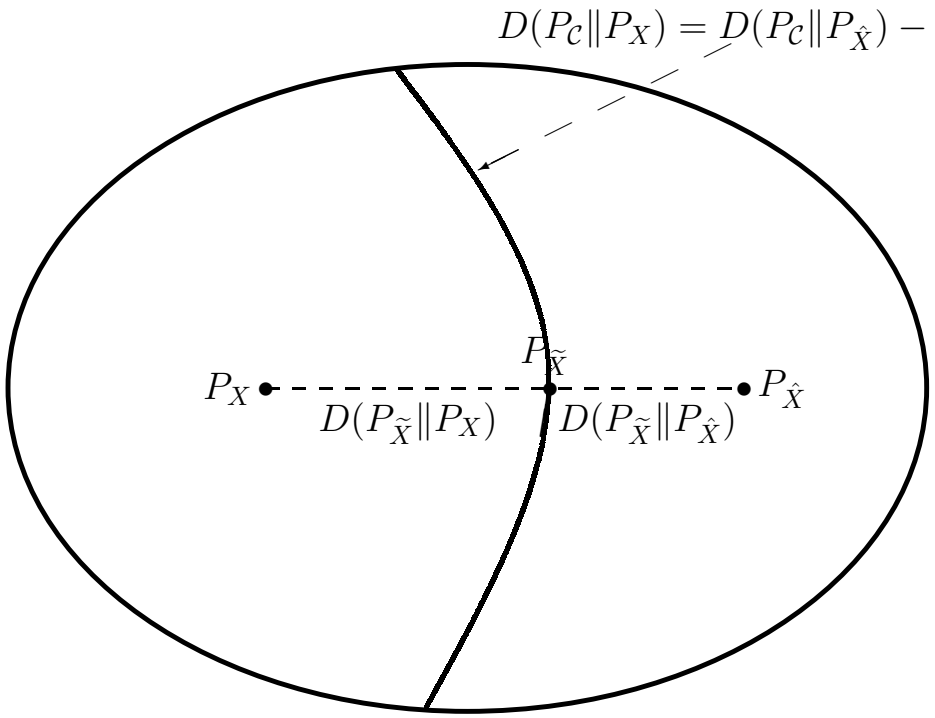
$$\log\frac{P_{\widetilde{X}}(x)}{P_{\hat{X}}(x)} + 1 + \lambda\log\frac{P_X(x)}{P_{\hat{X}}(x)} + \nu = 0.$$

  Solving these equations, we obtain the optimal $P_{\widetilde{X}}$ is of the form

$$P_{\widetilde{X}}(x) = P_\lambda(x) \stackrel{\triangle}{=} \frac{P_X^\lambda(x)P_{\hat{X}}^{1-\lambda}(x)}{\sum_{a\in\mathcal{X}} P_X^\lambda(a)P_{\hat{X}}^{1-\lambda}(a)}.$$

- The geometrical explanation for P_λ is that it locates on the "straight line" between P_X and $P_{\hat{X}}$ (in the sense of divergence measure) over the probability space.

Exponent of Bayesian cost



The divergence view on hypothesis testing.

Exponent of Bayesian cost

- When $\lambda \to 0$, $P_\lambda \to P_{\hat{X}}$; when $\lambda \to 1$, $P_\lambda \to P_X$.
- Usually, P_λ is named the *tilted* or *twisted* distribution.
- The value of λ is dependent on $\tau' = (1/n)\log\tau$.
- It is known from detection theory that the best τ for Bayes testing is π_1/π_0, which is fixed.
- Therefore,

$$\tau' = \lim_{n\to\infty} \frac{1}{n} \log \frac{\pi_1}{\pi_0} = 0,$$

  which implies that the optimal exponent for Bayes error is the minimum of $D(P_\lambda \| P_X)$ subject to $D(P_\lambda \| P_X) = D(P_\lambda \| P_{\hat{X}})$, namely the mid-point ($\lambda = 1/2$) of the line segment $(P_X, P_{\hat{X}})$ on probability space. This quantity is called the *Chernoff bound*.

Large deviations theory

- The large deviations theory basically consider the technique of computing the exponent of an exponentially decayed probability.

Tilted or twisted distribution

- Suppose the probability of a set $P_X(\mathcal{A}_n)$ decreasing down to zero exponentially fact, and its exponent is equal to $a > 0$.
- Over the probability space, let $\mathcal{P}$ denote the set of those distributions $P_{\widetilde{X}}$ satisfying $P_{\widetilde{X}}(\mathcal{A}_n)$ exhibits zero exponent.
- Then applying similar concept as Sanov's theorem, we can expect that

$$a = \min_{P_{\widetilde{X}} \in \mathcal{P}} D(P_{\widetilde{X}} \| P_X).$$

- Now suppose the minimizer of the above function happens at $f(P_{\widetilde{X}}) = \tau$ for some constant τ and some differentiable function $f(\cdot)$, the minimizer should be of the form

$$(\forall\ a \in \mathcal{X})\ P_{\widetilde{X}}(a) = \frac{P_X(a) e^{\lambda \frac{\partial f(P_{\widetilde{X}})}{\partial P_{\widetilde{X}}(a)}}}{\sum_{a' \in \mathcal{X}} P_X(a') e^{\lambda \frac{\partial f(P_{\widetilde{X}})}{\partial P_{\widetilde{X}}(a')}}}.$$

  As a result, $P_{\widetilde{X}}$ is the resultant distribution from P_X exponentially twisted via the partial derivative of the function f.
- Note that $P_{\widetilde{X}}$ is usually written as $P_X^{(\lambda)}$ since it is generated by twisting P_X with twisted factor λ.

Conventional twisted distribution

- The conventional definition for twisted distribution is based on the divergence function, i.e., $f(P_{\widetilde{X}}) = D(P_{\widetilde{X}} \| P_{\hat{X}}) - D(P_{\widetilde{X}} \| P_X)$.
- Since

$$\frac{\partial D(P_{\widetilde{X}} \| P_X)}{\partial P_{\widetilde{X}}(a)} = \log \frac{P_{\widetilde{X}}(a)}{P_X(a)} + 1,$$

the twisted distribution becomes

$$\begin{aligned}(\forall\ a \in \mathcal{X})\ P_{\widetilde{X}}(a) &= \frac{P_X(a) e^{\lambda \log \frac{P_{\hat{X}}(a)}{P_X(a)}}}{\sum_{a' \in \mathcal{X}} P_X(a') e^{\lambda \log \frac{P_{\hat{X}}(a')}{P_X(a')}}} \\ &= \frac{P_X^{1-\lambda}(a) P_{\hat{X}}^{\lambda}(a)}{\sum_{a' \in \mathcal{X}} P_X^{1-\alpha}(a) P_{\hat{X}}^{\lambda}(a)}\end{aligned}$$

Cramer's theorem

- **Question:** Consider a sequence of i.i.d. random variables, X^n, and suppose that we are interested in the probability of the set

$$\left\{ \frac{X_1 + \cdots + X_n}{n} > \tau \right\}.$$

- Observe that $(X_1 + \cdots + X_n)/n$ can be re-written as

$$\sum_{a \in \mathcal{X}} a \cdot P_{\mathcal{C}}(a).$$

- Therefore, the function f becomes

$$f(P_{\tilde{X}}) = \sum_{a \in \mathcal{X}} a P_{\tilde{X}}(a),$$

  and its partial derivative with respect to $P_{\tilde{X}}(a)$ is a.

- The resultant twisted distribution is

$$(\forall\ a \in \mathcal{X})\ P_X^{(\lambda)}(a) = \frac{P_X(a) e^{\lambda a}}{\sum_{a' \in \mathcal{X}} P_X(a') e^{\lambda a'}}.$$

- So the exponent of $P_{X^n}\{(X_1 + \cdots + X_n)/n > \tau\}$ is

$$\min_{\{P_{\tilde{X}}\ :\ D(P_{\tilde{X}} \| P_X) > \tau\}} D(P_{\tilde{X}} \| P_X) = \min_{\{P_X^{(\lambda)}\ :\ D(P_X^{(\lambda)} \| P_X) > \tau\}} D(P_X^{(\lambda)} \| P_X).$$

Cramer's theorem

- It should be pointed out that $\sum_{a' \in \mathcal{X}} P_X(a') e^{\lambda a'}$ is the moment generating function of P_X.
- The conventional Cramer's result does not use the divergence. Instead, it introduced the large deviation rate function, defined by

$$I_X(x) \stackrel{\triangle}{=} \sup_{\theta \in \Re} [\theta x - \log M_X(\theta)], \tag{8.2.4}$$

  where $M_X(\theta)$ is the moment generating function of X.
- Using his statement, the exponent of the above probability is respectively lower- and upper bounded by

$$\inf_{x \geq \tau} I_X(x) \quad \text{and} \quad \inf_{x > \tau} I_X(x).$$

  An example on how to obtain the exponent bounds is illustrated in the next subsection.

Exponent and moment generating function

A) Preliminaries : Observe that since $E[X] = \mu < \lambda$ and $E[|X - \mu|^2] < \infty$,

$$Pr\left\{\frac{X_1 + \cdots + X_n}{n} \geq \lambda\right\} \to 0 \text{ as } n \to \infty.$$

Hence, we can compute its rate of convergence (to zero).

B) Upper bound of the probability :

$$\begin{aligned}
& Pr\left\{\frac{X_1 + \cdots + X_n}{n} \geq \lambda\right\} \\
= \; & Pr\left\{\theta(X_1 + \cdots + X_n) \geq \theta n\lambda\right\}, \text{ for any } \theta > 0 \\
= \; & Pr\left\{\exp\left(\theta(X_1 + \cdots + X_n)\right) \geq \exp\left(\theta n\lambda\right)\right\} \\
\leq \; & \frac{E\left[\exp\left(\theta(X_1 + \cdots + X_n)\right)\right]}{\exp\left(\theta n\lambda\right)} \\
= \; & \frac{E^n\left[\exp\left(\theta X\right)\right]}{\exp\left(\theta n\lambda\right)} \\
= \; & \left(\frac{M_X(\theta)}{\exp\left(\theta\lambda\right)}\right)^n.
\end{aligned}$$

Hence,

$$\liminf_{n\to\infty} -\frac{1}{n} Pr\left\{\frac{X_1 + \cdots + X_n}{n} > \lambda\right\} \geq \theta\lambda - \log M_X(\theta).$$

Since the above inequality holds for every $\theta > 0$, we have

$$\begin{aligned}\liminf_{n\to\infty} -\frac{1}{n} Pr\left\{\frac{X_1 + \cdots + X_n}{n} > \lambda\right\} &\geq \max_{\theta>0}[\theta\lambda - \log M_X(\theta)] \\ &= \theta^*\lambda - \log M_X(\theta^*),\end{aligned}$$

where $\theta^* > 0$ is the optimizer of the maximum operation. (The positivity of θ^* can be easily verified by the concavity of the function $\theta\lambda - \log M_X(\theta)$ in θ, and it derivative at $\theta = 0$ equals $(\lambda - \mu)$ which is strictly greater than 0.) Consequently,

$$\begin{aligned}\liminf_{n\to\infty} -\frac{1}{n} Pr\left\{\frac{X_1 + \cdots + X_n}{n} > \lambda\right\} &\geq \theta^*\lambda - \log M_X(\theta^*) \\ &= \sup_{\theta\in\Re}[\theta\lambda - \log M_X(\theta)] = I_X(\lambda).\end{aligned}$$

C) Lower bound of the probability : omit.

Theories on Large deviations

In this section, we will derive inequalities on the exponent of the probability, $Pr\{Z_n/n \in [a,b]\}$, which is a slight extension of the Gärtner-Ellis theorem.

Definition 8.8 In this subsection, $\{Z_n\}_{n=1}^{\infty}$ will denote an infinite sequence of *arbitrary* random variables.

Definition 8.9 Define

$$\varphi_n(\theta) \stackrel{\triangle}{=} \frac{1}{n} \log E\left[\exp\left\{\theta Z_n\right\}\right] \quad \text{and} \quad \bar{\varphi}(\theta) \stackrel{\triangle}{=} \limsup_{n\to\infty} \varphi_n(\theta).$$

The *sup-large deviation rate function* of an arbitrary random sequence $\{Z_n\}_{n=1}^{\infty}$ is defined as

$$\bar{I}(x) \stackrel{\triangle}{=} \sup_{\{\theta\in\Re\,:\,\bar{\varphi}(\theta)>-\infty\}} \left[\theta x - \bar{\varphi}(\theta)\right]. \tag{8.3.5}$$

The range of the supremum operation in (8.3.5) is always non-empty since $\bar{\varphi}(0) = 0$, i.e. $\{\theta \in \Re \,:\, \bar{\varphi}(\theta) > -\infty\} \neq \emptyset$. Hence, $\bar{I}(x)$ is always defined. With the above definition, the first extension theorem of Gärtner-Ellis can be proposed as follows.

Theorem 8.10 For $a, b \in \Re$ and $a \leq b$,

$$\limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a,b]\right\} \leq -\inf_{x\in[a,b]} \bar{I}(x).$$

- The bound obtained in the above theorem is not in general tight.

Example 8.11 Suppose that $Pr\{Z_n = 0\} = 1 - e^{-2n}$, and $Pr\{Z_n = -2n\} = e^{-2n}$. Then from Definition 8.9, we have

$$\varphi_n(\theta) \stackrel{\triangle}{=} \frac{1}{n} \log E\left[e^{\theta Z_n}\right] = \frac{1}{n} \log\left[1 - e^{-2n} + e^{-(\theta+1)\cdot 2\cdot n}\right],$$

and

$$\bar{\varphi}(\theta) \stackrel{\triangle}{=} \limsup_{n\to\infty} \varphi_n(\theta) = \begin{cases} 0, & \text{for } \theta \geq -1; \\ -2(\theta+1), & \text{for } \theta < -1. \end{cases}$$

Hence, $\{\theta \in \Re : \bar{\varphi}(\theta) > -\infty\} = \Re$ and

$$\begin{aligned} \bar{I}(x) &= \sup_{\theta\in\Re}[\theta x - \bar{\varphi}(\theta)] \\ &= \sup_{\theta\in\Re}[\theta x + 2(\theta+1)\mathbf{1}\{\theta < -1)\}] \\ &= \begin{cases} -x, & \text{for } -2 \leq x \leq 0; \\ \infty, & \text{otherwise,} \end{cases} \end{aligned}$$

where $\mathbf{1}\{\cdot\}$ represents the indicator function of a set.

Consequently, by Theorem 8.10,

$$\limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a,b]\right\} \leq -\inf_{x\in[a,b]} \bar{I}(x)$$
$$= \begin{cases} 0, & \text{for } 0 \in [a,b]; \\ b, & \text{for } b \in [-2,0]; \\ -\infty, & \textit{otherwise}. \end{cases}$$

The exponent of $Pr\{Z_n/n \in [a,b]\}$ in the above example is indeed given by

$$\lim_{n\to\infty} \frac{1}{n} \log P_{Z^n}\left\{\frac{Z_n}{n} \in [a,b]\right\} = -\inf_{x\in[a,b]} I^*(x),$$

where

$$I^*(x) = \begin{cases} 2, & \text{for } x = -2; \\ 0, & \text{for } x = 0; \\ \infty, & \text{otherwise.} \end{cases} \tag{8.3.6}$$

Thus, the upper bound obtained in Theorem 8.10 is not tight.

Definition 8.12 Define

$$\varphi_n(\theta;h)\stackrel{\triangle}{=}\frac{1}{n}\log E\left[\exp\left\{n\cdot\theta\cdot h\left(\frac{Z_n}{n}\right)\right\}\right] \quad \text{and} \quad \bar{\varphi}_h(\theta)\stackrel{\triangle}{=}\limsup_{n\to\infty}\varphi_n(\theta;h),$$

where $h(\cdot)$ is a given real-valued continuous function. The *twisted sup-large deviation rate function* of an arbitrary random sequence $\{Z_n\}_{n=1}^{\infty}$ with respect to a real-valued continuous function $h(\cdot)$ is defined as

$$\bar{J}_h(x)\stackrel{\triangle}{=}\sup_{\{\theta\in\Re\,:\,\bar{\varphi}_h(\theta)>-\infty\}}\left[\theta\cdot h(x)-\bar{\varphi}_h(\theta)\right]. \tag{8.3.7}$$

Theorem 8.13 Suppose that $h(\cdot)$ is a real-valued continuous function. Then for $a,b\in\Re$ and $a\le b$,

$$\limsup_{n\to\infty}\frac{1}{n}\log Pr\left\{\frac{Z_n}{n}\in[a,b]\right\}\le-\inf_{x\in[a,b]}\bar{J}_h(x).$$

Example 8.14 Let us, again, investigate the $\{Z_n\}_{n=1}^{\infty}$ defined in Example 8.11. Take

$$h(x) = \frac{1}{2}(x+2)^2 - 1.$$

Then from Definition 8.12, we have

$$\begin{aligned}\varphi_n(\theta; h) &\stackrel{\triangle}{=} \frac{1}{n}\log E\left[\exp\left\{n\theta h(Z_n/n)\right\}\right] \\ &= \frac{1}{n}\log\left[\exp\left\{n\theta\right\} - \exp\left\{n(\theta-2)\right\} + \exp\left\{-n(\theta+2)\right\}\right],\end{aligned}$$

and

$$\bar{\varphi}_h(\theta) \stackrel{\triangle}{=} \limsup_{n\to\infty} \varphi_n(\theta; h) = \begin{cases} -(\theta+2), & \text{for } \theta \le -1; \\ \theta, & \text{for } \theta > -1. \end{cases}$$

Hence, $\{\theta \in \Re \ : \ \bar{\varphi}_h(\theta) > -\infty\} = \Re$ and

$$\bar{J}_h(x) \stackrel{\triangle}{=} \sup_{\theta\in\Re}\left[\theta h(x) - \bar{\varphi}_h(\theta)\right] = \begin{cases} -\frac{1}{2}(x+2)^2 + 2, & \text{for } x \in [-4, 0]; \\ \infty, & \text{otherwise.} \end{cases}$$

Consequently, by Theorem 8.13,

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a,b]\right\} \\
\leq\ & -\inf_{x\in[a,b]} \bar{J}_h(x) \\
=\ & \begin{cases} -\min\left\{-\dfrac{(a+2)^2}{2}, -\dfrac{(b+2)^2}{2}\right\} - 2, & \text{for } -4 \leq a < b \leq 0; \\ 0, & \text{for } a > 0 \text{ or } b < -4; \\ -\infty, & \text{otherwise.} \end{cases}
\end{aligned} \tag{8.3.8}$$

For $b \in (-2,0)$ and $a \in \left[-2-\sqrt{2b-4}, b\right)$, the upper bound attained in the previous example is strictly less than that given in Example 8.11, and hence, an improvement is obtained. However, for $b \in (-2,0)$ and $a < -2-\sqrt{2b-4}$, the upper bound in (8.3.8) is actually looser. Accordingly, we combine the two upper bounds from Examples 8.11 and 8.14 to get

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a,b]\right\} \ &\leq\ -\max\left\{\inf_{x\in[a,b]} \bar{J}_h(x), \inf_{x\in[a,b]} \bar{I}(x)\right\} \\
&= \begin{cases} 0, & \text{for } 0 \in [a,b]; \\ \dfrac{1}{2}(b+2)^2 - 2, & \text{for } b \in [-2,0]; \\ -\infty, & \text{otherwise.} \end{cases}
\end{aligned}$$

Theorem 8.15 For $a, b \in \Re$ and $a \leq b$,

$$\limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a,b]\right\} \leq -\inf_{x\in[a,b]} \bar{J}(x),$$

where $\bar{J}(x) \stackrel{\triangle}{=} \sup_{h\in\mathcal{H}} \bar{J}_h(x)$ and $\mathcal{H}$ is the set of all real-valued continuous functions.

Example 8.16 Let us again study the $\{Z_n\}_{n=1}^{\infty}$ in Example 8.11 (also in Example 8.14). Suppose $c > 1$. Take $h_c(x) = c_1(x + c_2)^2 - c$, where

$$c_1 \stackrel{\triangle}{=} \frac{c + \sqrt{c^2 - 1}}{2} \quad \text{and} \quad c_2 \stackrel{\triangle}{=} \frac{2\sqrt{c+1}}{\sqrt{c+1} + \sqrt{c-1}}.$$

Then from Definition 8.12, we have

$$\begin{aligned} \varphi_n(\theta; h_c) &\stackrel{\triangle}{=} \frac{1}{n} \log E\left[\exp\left\{n\theta h_c\left(\frac{Z_n}{n}\right)\right\}\right] \\ &= \frac{1}{n} \log\left[\exp\{n\theta\} - \exp\{n(\theta-2)\} + \exp\{-n(\theta+2)\}\right], \end{aligned}$$

and

$$\bar{\varphi}_{h_c}(\theta) \stackrel{\triangle}{=} \limsup_{n\to\infty} \varphi_n(\theta; h_c) = \begin{cases} -(\theta+2), & \text{for } \theta \leq -1; \\ \theta, & \text{for } \theta > -1. \end{cases}$$

Extension of Gärtner-Ellis upper bounds

Hence, $\{\theta \in \Re \ : \ \bar{\varphi}_{h_c}(\theta) > -\infty\} = \Re$ and

$$\begin{aligned} \bar{J}_{h_c}(x) &= \sup_{\theta \in \Re} [\theta h_c(x) - \bar{\varphi}_{h_c}(\theta)] \\ &= \begin{cases} -c_1(x+c_2)^2 + c + 1, & \text{for } x \in [-2c_2, 0]; \\ \infty, & \text{otherwise.} \end{cases} \end{aligned}$$

From Theorem 8.15,

$$\bar{J}(x) = \sup_{h \in \mathcal{H}} \bar{J}_h(x) \geq \max\{\liminf_{c \to \infty} \bar{J}_{h_c}(x), \bar{I}(x)\} = I^*(x),$$

where $I^*(x)$ is defined in (8.3.6). Consequently,

$$\begin{aligned} \limsup_{n \to \infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a,b]\right\} &\leq -\inf_{x \in [a,b]} \bar{J}(x) \\ &\leq -\inf_{x \in [a,b]} I^*(x) \\ &= \begin{cases} 0, & \text{if } 0 \in [a,b]; \\ -2, & \text{if } -2 \in [a,b] \text{ and } 0 \notin [a,b]; \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

and a tight upper bound is finally obtained!

Extension of Gärtner-Ellis upper bounds

Definition 8.17 Define $\underline{\varphi}_h(\theta) \stackrel{\triangle}{=} \liminf_{n\to\infty} \varphi_n(\theta; h)$, where $\varphi_n(\theta; h)$ was defined in Definition 8.12. The *twisted inf-large deviation rate function* of an arbitrary random sequence $\{Z_n\}_{n=1}^{\infty}$ with respect to a real-valued continuous function $h(\cdot)$ is defined as

$$\underline{J}_h(x) \stackrel{\triangle}{=} \sup_{\{\theta\in\Re \,:\, \underline{\varphi}_h(\theta) > -\infty\}} \left[\theta \cdot h(x) - \underline{\varphi}_h(\theta)\right].$$

Theorem 8.18 For $a, b \in \Re$ and $a \leq b$,

$$\liminf_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a, b]\right\} \leq - \inf_{x\in[a,b]} \underline{J}(x),$$

where $\underline{J}(x) \stackrel{\triangle}{=} \sup_{h\in\mathcal{H}} \underline{J}_h(x)$ and $\mathcal{H}$ is the set of all real-valued continuous functions.

- **Hope to know when**

$$\limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{ \frac{Z_n}{n} \in (a,b) \right\} \geq - \inf_{x\in(a,b)} \bar{J}_h(x). \tag{8.3.9}$$

Definition 8.19 Define the *sup-Gärtner-Ellis set* with respect to a real-valued continuous function $h(\cdot)$ as

$$\bar{\mathcal{D}}_h \stackrel{\triangle}{=} \bigcup_{\{\theta\in\Re \,:\, \bar{\varphi}_h(\theta)>-\infty\}} \bar{\mathcal{D}}(\theta;h)$$

where

$$\bar{\mathcal{D}}(\theta;h) \stackrel{\triangle}{=} \left\{ x\in\Re \,:\, \limsup_{t\downarrow 0} \frac{\bar{\varphi}_h(\theta+t)-\bar{\varphi}_h(\theta)}{t} \right.$$
$$\left. \leq h(x) \leq \liminf_{t\downarrow 0} \frac{\bar{\varphi}_h(\theta)-\bar{\varphi}_h(\theta-t)}{t} \right\}.$$

Let us briefly remark on the *sup-Gärtner-Ellis set* defined above.

- It can be derived that the *sup-Gärtner-Ellis set* is reduced to

$$\bar{\mathcal{D}}_h \stackrel{\triangle}{=} \bigcup_{\{\theta\in\Re \,:\, \bar{\varphi}_h(\theta)>-\infty\}} \{x\in\Re \,:\, \bar{\varphi}_h'(\theta)=h(x)\},$$

  if the derivative $\bar{\varphi}_h'(\theta)$ exists for all θ.

Theorem 8.20 Suppose that $h(\cdot)$ is a real-valued continuous function. Then if $(a,b) \subset \bar{\mathcal{D}}_h$,

$$\limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in (a,b)\right\} \geq - \inf_{x\in(a,b)} \bar{J}_h(x).$$

Example 8.21 Suppose $Z_n = X_1 + \cdots + X_n$, where $\{X_i\}_{i=1}^n$ are i.i.d. Gaussian random variables with mean 1 and variance 1 if n is even, and with mean -1 and variance 1 if n is odd. Then the exact large deviation rate formula $\bar{I}^*(x)$ that satisfies for all $a < b$,

$$\begin{aligned} - \inf_{x\in[a,b]} \bar{I}^*(x) &\geq \limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in [a,b]\right\} \\ &\geq \limsup_{n\to\infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in (a,b)\right\} \geq - \inf_{x\in(a,b)} \bar{I}^*(x) \end{aligned}$$

is

$$\bar{I}^*(x) = \frac{(|x|-1)^2}{2}. \tag{8.3.10}$$

Case A: $h(x) = x$.

For the affine $h(\cdot)$, $\varphi_n(\theta) = \theta + \theta^2/2$ when n is even, and $\varphi_n(\theta) = -\theta + \theta^2/2$

when n is odd. Hence, $\bar{\varphi}(\theta) = |\theta| + \theta^2/2$, and

$$\begin{aligned}\bar{\mathcal{D}}_h &= \left(\bigcup_{\theta>0}\{v \in \Re \ : \ v = 1+\theta\}\right) \bigcup \left(\bigcup_{\theta<0}\{v \in \Re \ : \ v = -1+\theta\}\right) \\ &\quad \bigcup \{v \in \Re \ : \ 1 \le v \le -1\} \\ &= (1,\infty) \cup (-\infty, -1).\end{aligned}$$

Therefore, Theorem 8.20 cannot be applied to any a and b with $(a,b) \cap [-1,1] \neq \emptyset$.

By deriving

$$\bar{I}(x) = \sup_{\theta \in \Re}\{x\theta - \bar{\varphi}(\theta)\} = \begin{cases} \dfrac{(|x|-1)^2}{2}, & \text{for } |x| > 1; \\ 0, & \text{for } |x| \le 1, \end{cases}$$

we obtain for any $a \in (-\infty, 1) \cup (1, \infty)$,

$$\begin{aligned}&\lim_{\varepsilon \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \log Pr\left\{\frac{Z_n}{n} \in (a-\varepsilon, a+\varepsilon)\right\} \\ \ge\ & -\lim_{\varepsilon \downarrow 0} \inf_{x \in (a-\varepsilon, a+\varepsilon)} \bar{I}(x) = -\frac{(|a|-1)^2}{2},\end{aligned}$$

which can be shown tight by Theorem 8.13 (or directly by (8.3.10)). Note that the above inequality does not hold for any $a \in (-1, 1)$. To fill the gap, a different $h(\cdot)$ must be employed.

Case B: $h(x) = |x|$.

For n even,

$$\begin{aligned}
& E\left[e^{n\theta h(Z_n/n)}\right] \\
= \; & E\left[e^{n\theta|Z_n/n-a|}\right] \\
= \; & \int_{-\infty}^{na} e^{-\theta x+n\theta a}\frac{1}{\sqrt{2\pi n}}e^{-(x-n)^2/(2n)}dx + \int_{na}^{\infty} e^{\theta x-n\theta a}\frac{1}{\sqrt{2\pi n}}e^{-(x-n)^2/(2n)}dx \\
= \; & e^{n\theta(\theta-2+2a)/2}\int_{-\infty}^{na}\frac{1}{\sqrt{2\pi n}}e^{-[x-n(1-\theta)]^2/(2n)}dx \\
& + e^{n\theta(\theta+2-2a)/2}\int_{na}^{\infty}\frac{1}{\sqrt{2\pi n}}e^{-[x-n(1+\theta)]^2/(2n)}dx \\
= \; & e^{n\theta(\theta-2+2a)/2}\cdot\Phi\left((\theta+a-1)\sqrt{n}\right) + e^{n\theta(\theta+2-2a)/2}\cdot\Phi\left((\theta-a+1)\sqrt{n}\right),
\end{aligned}$$

where $\Phi(\cdot)$ represents the unit Gaussian cdf.

Similarly, for n odd,

$$\begin{aligned} & E\left[e^{n\theta h(Z_n/n)}\right] \\ = \; & e^{n\theta(\theta+2+2a)/2} \cdot \Phi\left((\theta+a+1)\sqrt{n}\right) + e^{n\theta(\theta-2-2a)/2} \cdot \Phi\left((\theta-a-1)\sqrt{n}\right). \end{aligned}$$

Observe that for any $b \in \Re$,

$$\lim_{n\to\infty} \frac{1}{n} \log \Phi(b\sqrt{n}) = \begin{cases} 0, & \text{for } b \geq 0; \\ -\dfrac{b^2}{2}, & \text{for } b < 0. \end{cases}$$

Hence,

$$\bar{\varphi}_h(\theta) = \begin{cases} -\dfrac{(|a|-1)^2}{2}, & \text{for } \theta < |a|-1; \\ \dfrac{\theta[\theta+2(1-|a|)]}{2}, & \text{for } |a|-1 \leq \theta < 0; \\ \dfrac{\theta[\theta+2(1+|a|)]}{2}, & \text{for } \theta \geq 0. \end{cases}$$

Therefore,

$$\begin{aligned}\bar{\mathcal{D}}_h &= \left(\bigcup_{\theta>0}\{x\in\Re\ :\ |x-a|=\theta+1+|a|\}\right)\\ &\qquad \bigcup\left(\bigcup_{\theta<0}\{x\in\Re\ :\ |x-a|=\theta+1-|a|\}\right)\\ &= (-\infty, a-1-|a|)\cup(a-1+|a|, a+1-|a|)\cup(a+1+|a|,\infty)\end{aligned}$$

and

$$\bar{J}_h(x)=\begin{cases}\dfrac{(|x-a|-1+|a|)^2}{2}, & \text{for } a-1+|a|<x<a+1-|a|;\\ \dfrac{(|x-a|-1-|a|)^2}{2}, & \text{for } x>a+1+|a| \text{ or } x<a-1-|a|;\\ 0, & \text{otherwise.}\end{cases}$$

(8.3.11)

We then apply Theorem 8.20 to obtain

$$\begin{aligned} & \lim_{\varepsilon \downarrow 0} \limsup_{n \to \infty} \frac{1}{n} \log Pr \left\{ \frac{Z_n}{n} \in (a-\varepsilon, a+\varepsilon) \right\} \\ \geq & -\lim_{\varepsilon \downarrow 0} \inf_{x \in (a-\varepsilon, a+\varepsilon)} \bar{J}_h(x) \\ = & -\lim_{\varepsilon \downarrow 0} \frac{(\varepsilon - 1 + |a|)^2}{2} = -\frac{(|a|-1)^2}{2}. \end{aligned}$$

Note that the above lower bound is valid for any $a \in (-1, 1)$, and can be shown tight, again, by Theorem 8.13 (or directly by (8.3.10)).

Finally, by combining the results of Cases A) and B), the true large deviation rate of $\{Z_n\}_{n \geq 1}$ is completely characterized.

Definition 8.22 Define the *inf-Gärtner-Ellis set* with respect to a real-valued continuous function $h(\cdot)$ as

$$\underline{\mathcal{D}}_h \stackrel{\triangle}{=} \bigcup_{\{\theta \in \Re \, : \, \underline{\varphi}_h(\theta) > -\infty\}} \underline{\mathcal{D}}(\theta; h)$$

where

$$\underline{\mathcal{D}}(\theta; h) \stackrel{\triangle}{=} \left\{ x \in \Re \, : \limsup_{t \downarrow 0} \frac{\underline{\varphi}_h(\theta + t) - \underline{\varphi}_h(\theta)}{t} \right.$$
$$\left. \leq h(x) \leq \liminf_{t \downarrow 0} \frac{\underline{\varphi}_h(\theta) - \underline{\varphi}_h(\theta - t)}{t} \right\}.$$

Theorem 8.23 Suppose that $h(\cdot)$ is a real-valued continuous function. Then if $(a, b) \subset \underline{\mathcal{D}}_h$,

$$\liminf_{n \to \infty} \frac{1}{n} \log Pr \left\{ \frac{Z_n}{n} \in (a, b) \right\} \geq - \inf_{x \in (a,b)} \underline{J}_h(x).$$

Properties

Property 8.24 Let $\bar{I}(x)$ and $\underline{I}(x)$ be the sup- and inf- large deviation rate functions of an infinite sequence of arbitrary random variables $\{Z_n\}_{n=1}^{\infty}$, respectively. Denote $m_n = (1/n)E[Z_n]$. Let $\bar{m} \stackrel{\triangle}{=} \limsup_{n\to\infty} m_n$ and $\underline{m} \stackrel{\triangle}{=} \liminf_{n\to\infty} m_n$. Then

1. $\bar{I}(x)$ and $\underline{I}(x)$ are both convex.
2. $\bar{I}(x)$ is continuous over $\{x \in \Re \ : \ \bar{I}(x) < \infty\}$. Likewise, $\underline{I}(x)$ is continuous over $\{x \in \Re \ : \ \underline{I}(x) < \infty\}$.
3. $\bar{I}(x)$ gives its minimum value 0 at $\underline{m} \leq x \leq \bar{m}$.
4. $\underline{I}(x) \geq 0$. But $\underline{I}(x)$ does not necessary give its minimum value at both $x = \bar{m}$ and $x = \underline{m}$.

Properties

Property 8.25 Suppose that $h(\cdot)$ is a real-valued continuous function. Let $\bar{J}_h(x)$ and $\underline{J}_h(x)$ be the corresponding twisted sup- and inf- large deviation rate functions, respectively. Denote $m_n(h) \stackrel{\triangle}{=} E[h(Z_n/n)]$. Let

$$\bar{m}_h \stackrel{\triangle}{=} \limsup_{n\to\infty} m_n(h) \quad \text{and} \quad \underline{m}_h \stackrel{\triangle}{=} \liminf_{n\to\infty} m_n(h).$$

Then

1. $\bar{J}_h(x) \geq 0$, with equality holds if $\underline{m}_h \leq h(x) \leq \bar{m}_h$.
2. $\underline{J}_h(x) \geq 0$, but $\underline{J}_h(x)$ does not necessary give its minimum value at both $x = \bar{m}_h$ and $x = \underline{m}_h$.

Probabilitic subexponential behavior

- **subexponential behavior.**

  $a_n = (1/n)\exp\{-2n\}$ and $b_n = (1/\sqrt{n})\exp\{-2n\}$ have the same exponent, but contain different subexponential terms

Berry-Esseen theorem for compound i.i.d. sequence

- Berry-Esseen theorem states that the distribution of the sum of independent zero-mean random variables $\{X_i\}_{i=1}^n$, normalized by the standard deviation of the sum, differs from the Gaussian distribution by at most $C\, r_n/s_n^3$, where s_n^2 and r_n are respectively sums of the marginal variances and the marginal absolute third moments, and C is an absolute constant.
- Specifically, for every $a \in \Re$,

$$\left| Pr\left\{ \frac{1}{s_n}(X_1 + \cdots + X_n) \le a \right\} - \Phi(a) \right| \le C\, \frac{r_n}{s_n^3}, \tag{8.4.12}$$

  where $\Phi(\cdot)$ represents the unit Gaussian cdf.
- The striking feature of this theorem is that the upper bound depends only on the variance and the absolute third moment, and hence, can provide a good asymptotic estimate based on only the first three moments.
- The absolute constant C is commonly 6. When $\{X_n\}_{i=1}^n$ are identically distributed, in addition to independent, the absolute constant can be reduced to 3, and has been reported to be improved down to 2.05.

Definition: compound i.i.d. sequence. The samples that we concern in this section actually consists of two i.i.d. sequences (and, is therefore named *compound i.i.d. sequence*.)

Lemma 8.26 (smoothing lemma) Fix the bandlimited filtering function

$$v_{_T}(x) \stackrel{\triangle}{=} \frac{1-\cos(Tx)}{\pi T x^2} = \frac{2\sin^2(Tx/2)}{\pi T x^2} = \frac{T}{2\pi}\mathrm{sinc}^2\left(\frac{Tx}{2\pi}\right) = Four^{-1}\left[\Lambda\left(\frac{f}{T/(2\pi)}\right)\right].$$

For any cumulative distribution function $H(\cdot)$ on the real line $\Re$,

$$\sup_{x\in\Re}|\Delta_T(x)| \geq \frac{1}{2}\,\eta - \frac{6}{T\pi\sqrt{2\pi}}\,h\left(\frac{T\sqrt{2\pi}}{2}\,\eta\right),$$

where

$$\Delta_T(t) \stackrel{\triangle}{=} \int_{-\infty}^{\infty}[H(t-x)-\Phi(t-x)]\times v_{_T}(x)dx, \quad \eta \stackrel{\triangle}{=} \sup_{x\in\Re}|H(x)-\Phi(x)|,$$

and

$$h(u) \stackrel{\triangle}{=} \begin{cases} u\displaystyle\int_u^{\infty}\frac{1-\cos(x)}{x^2}dx = \frac{\pi}{2}u + 1 - \cos(u) - u\int_0^u\frac{\sin(x)}{x}dx, & \text{if } u\geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Lemma 8.27 For any cumulative distribution function $H(\cdot)$ with characteristic function $\varphi_{_H}(\zeta)$,

$$\eta \leq \frac{1}{\pi}\int_{-T}^{T}\left|\varphi_{_H}(\zeta) - e^{-(1/2)\zeta^2}\right|\frac{d\zeta}{|\zeta|} + \frac{12}{T\pi\sqrt{2\pi}}\,h\left(\frac{T\sqrt{2\pi}}{2}\,\eta\right),$$

where η and $h(\cdot)$ are defined in Lemma 8.26.

Theorem 8.28 (BE theorem for compound i.i.d. sequences) Let $Y_n = \sum_{i=1}^{n} X_i$ be the sum of independent random variables, among which $\{X_i\}_{i=1}^{d}$ are identically Gaussian distributed, and $\{X_i\}_{i=d+1}^{n}$ are identically distributed but not necessarily Gaussian.

- Denote the mean-variance pair of X_1 and X_{d+1} by (μ, σ^2) and $(\hat{\mu}, \hat{\sigma}^2)$, respectively.
- Define
$$\rho \stackrel{\triangle}{=} E\left[|X_1 - \mu|^3\right], \quad \hat{\rho} \stackrel{\triangle}{=} E\left[|X_{d+1} - \hat{\mu}|^3\right]$$
and
$$s_n^2 = \mathrm{Var}[Y_n] = \sigma^2 d + \hat{\sigma}^2(n-d).$$
- Also denote the cdf of $(Y_n - E[Y_n])/s_n$ by $H_n(\cdot)$.

Then for all $y \in \Re$,

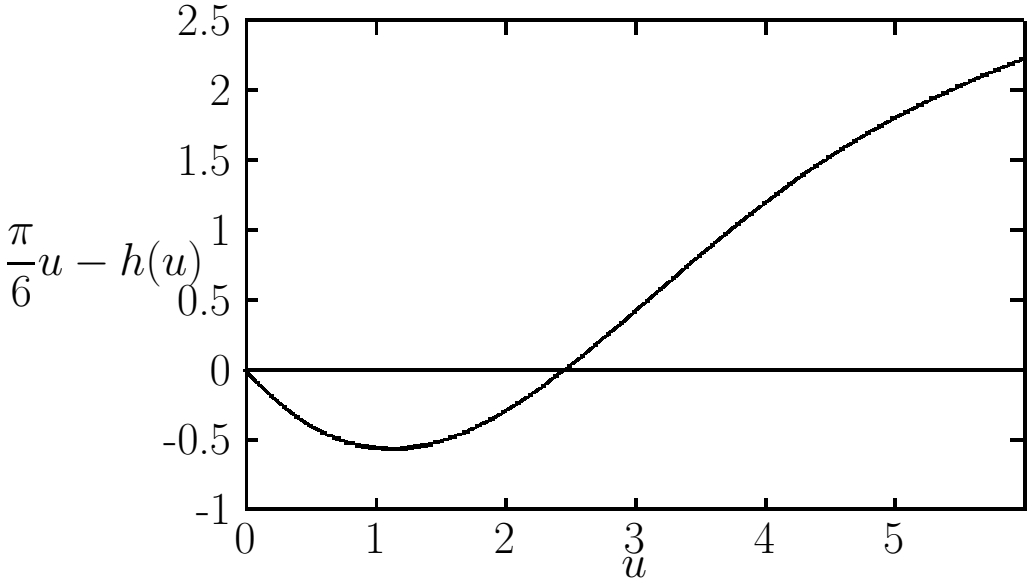
$$|H_n(y) - \Phi(y)| \leq C_{n,d} \frac{2}{\sqrt{\pi}} \frac{(n-d-1)}{\left(2(n-d) - 3\sqrt{2}\right)} \frac{\hat{\rho}}{\hat{\sigma}^2 s_n},$$

where $C_{n,d}$ is the unique positive number satisfying

$$\begin{aligned} & \frac{\pi}{6} C_{n,d} - h(C_{n,d}) \\ = \; & \frac{\sqrt{\pi}\left(2(n-d) - 3\sqrt{2}\right)}{12(n-d-1)} \left(\frac{\sqrt{6\pi}}{2\left(3-\sqrt{2}\right)^{3/2}} + \frac{9}{2(11-6\sqrt{2})\sqrt{n-d}} \right), \end{aligned}$$

provided that $n - d \geq 3$.

Berry-Esseen theorem for compound i.i.d. sequence



Function of $(\pi/6)u - h(u)$.

By letting $d = 0$, the Berry-Esseen inequality for i.i.d. sequences can also be readily obtained from the previous Theorem.

Corollary 8.29 (Berry-Esseen theorem for i.i.d. sequence) Let

$$Y_n = \sum_{i=1}^{n} X_i$$

be the sum of independent random variables with common marginal distribution. Denote the marginal mean and variance by $(\hat{\mu}, \hat{\sigma}^2)$. Define $\hat{\rho} \stackrel{\triangle}{=} E\left[|X_1 - \hat{\mu}|^3\right]$. Also denote the cdf of $(Y_n - n\hat{\mu})/(\sqrt{n}\hat{\sigma})$ by $H_n(\cdot)$. Then for all $y \in \Re$,

$$|H_n(y) - \Phi(y)| \leq C_n \frac{2(n-1)}{\sqrt{\pi}\left(2n - 3\sqrt{2}\right)} \frac{\hat{\rho}}{\hat{\sigma}^3\sqrt{n}},$$

where C_n is the unique positive solution of

$$\frac{\pi}{6}u - h(u) = \frac{\sqrt{\pi}\left(2n - 3\sqrt{2}\right)}{12(n-1)} \left(\frac{\sqrt{6\pi}}{2\left(3 - \sqrt{2}\right)^{3/2}} + \frac{9}{2(11 - 6\sqrt{2})\sqrt{n}} \right),$$

provided that $n \geq 3$.

- Let us briefly remark on the previous corollary. We observe from numericals that the quantity

$$C_n \frac{2}{\sqrt{\pi}} \frac{(n-1)}{\left(2n - 3\sqrt{2}\right)}$$

is decreasing in n, and ranges from 3.628 to 1.627 (cf. The picture in slide II:8-62.)

  - We can upperbound C_n by the unique positive solution D_n of

$$\frac{\pi}{6}u - h(u) = \frac{\sqrt{\pi}}{6}\left(\frac{\sqrt{6\pi}}{2\left(3-\sqrt{2}\right)^{3/2}} + \frac{9}{2(11-6\sqrt{2})\sqrt{n}}\right),$$

which is strictly decreasing in n. Hence,

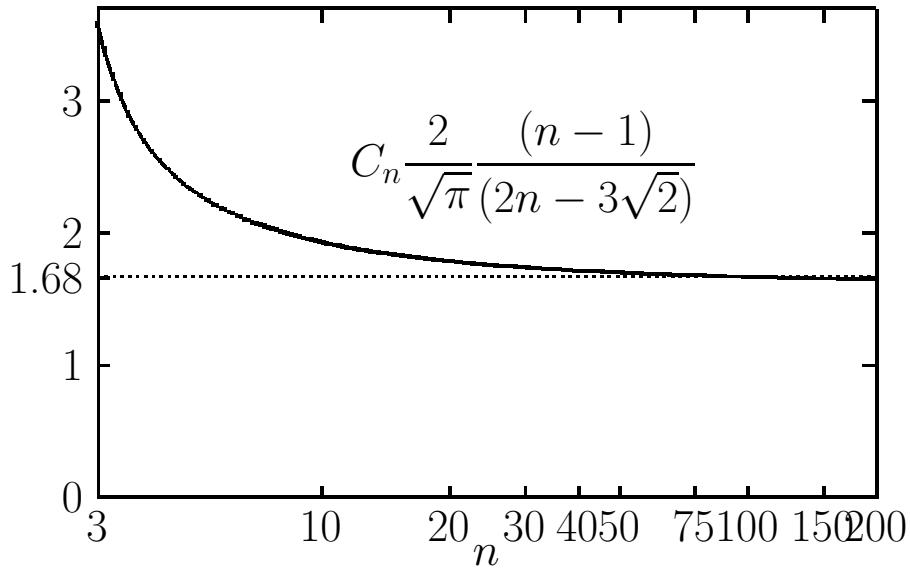
$$C_n \frac{2}{\sqrt{\pi}} \frac{(n-1)}{\left(2n - 3\sqrt{2}\right)} \leq E_n \stackrel{\triangle}{=} D_n \frac{2}{\sqrt{\pi}} \frac{(n-1)}{\left(2n - 3\sqrt{2}\right)},$$

and the right-hand-side of the above inequality is strictly decreasing (since both D_n and $(n-1)/(2n-3\sqrt{2})$ are decreasing) in n, and ranges from $E_3 = 4.1911, \ldots, E_9 = 2.0363, \ldots, E_{100} = 1.6833$ to $E_\infty = 1.6266$. If the property of strictly decreasingness is preferred, one can use D_n instead of C_n in the Berry-Esseen inequality. Note that both C_n and D_n converges to $2.8831\ldots$ as n goes to infinity.

Berry-Esseen theorem for compound i.i.d. sequence

- Numerical result shows that it lies below 2 when $n \geq 9$, and is smaller than 1.68 as $n \geq 100$. In other words, we can upperbound this quantity by 1.68 as $n \geq 100$, and therefore, establish a better estimate of the original Berry-Esseen constant.

Berry-Esseen theorem for compound i.i.d. sequence



The Berry-Esseen constant as a function of the sample size n. The sample size n is plotted in log-scale.

The general expression of the Neyman-Pearson type-II error exponent subject to a constant bound on the type-I error has been proved for arbitrary observations. In this section, we will state the results in terms of the ε-inf/sup-divergence rates.

Theorem 8.30 (Neyman-Pearson type-II error exponent for a fixed test level) Consider a sequence of random observations which is assumed to have a probability distribution governed by either $P_{\boldsymbol{X}}$ (null hypothesis) or $P_{\hat{\boldsymbol{X}}}$ (alternative hypothesis). Then, the type-II error exponent satisfies

$$\lim_{\delta \uparrow \delta} \bar{D}_\delta(\boldsymbol{X} \| \hat{\boldsymbol{X}}) \;\leq\; \limsup_{n \to \infty} -\frac{1}{n} \log \beta_n^*(\varepsilon) \;\leq\; \bar{D}_\varepsilon(\boldsymbol{X} \| \hat{\boldsymbol{X}})$$

$$\lim_{\delta \uparrow \varepsilon} \underline{D}_\delta(\boldsymbol{X} \| \hat{\boldsymbol{X}}) \;\leq\; \liminf_{n \to \infty} -\frac{1}{n} \log \beta_n^*(\varepsilon) \;\leq\; \underline{D}_\varepsilon(\boldsymbol{X} \| \hat{\boldsymbol{X}})$$

where $\beta_n^*(\varepsilon)$ represents the minimum type-II error probability subject to a fixed type-I error bound $\varepsilon \in [0, 1)$.

The general formula for Neyman-Pearson type-II error exponent subject to an exponential test level has also been proved in terms of the ε-inf/sup-divergence rates.

Theorem 8.31 (Neyman-Pearson type-II error exponent for an exponential test level) Fix $s \in (0,1)$ and $\varepsilon \in [0,1)$. It is possible to choose decision regions for a binary hypothesis testing problem with arbitrary datawords of blocklength n, (which are governed by either the null hypothesis distribution $P_{\boldsymbol{X}}$ or the alternative hypothesis distribution $P_{\hat{\boldsymbol{X}}}$,) such that

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n \;\geq\; \bar{D}_\varepsilon(\hat{\boldsymbol{X}}^{(s)}\|\hat{\boldsymbol{X}}) \text{ and } \limsup_{n\to\infty} -\frac{1}{n}\log\alpha_n \;\geq\; \underline{D}_{(1-\varepsilon)}(\hat{\boldsymbol{X}}^{(s)}\|\boldsymbol{X}), \tag{8.5.13}$$

or

$$\liminf_{n\to\infty} -\frac{1}{n}\log\beta_n \;\geq\; \underline{D}_\varepsilon(\hat{\boldsymbol{X}}^{(s)}\|\hat{\boldsymbol{X}}) \text{ and } \limsup_{n\to\infty} -\frac{1}{n}\log\alpha_n \;\geq\; \bar{D}_{(1-\varepsilon)}(\hat{\boldsymbol{X}}^{(s)}\|\boldsymbol{X}), \tag{8.5.14}$$

where $\hat{\boldsymbol{X}}^{(s)}$ exhibits the tilted distributions $\left\{P^{(s)}_{\hat{X}^n}\right\}_{n=1}^{\infty}$ defined by

$$dP^{(s)}_{\hat{X}^n}(x^n) \stackrel{\triangle}{=} \frac{1}{\Omega_n(s)} \exp\left\{ s\log\frac{dP_{X^n}}{dP_{\hat{X}^n}}(x^n)\right\} dP_{\hat{X}^n}(x^n),$$

and

$$\Omega_n(s) \stackrel{\triangle}{=} \int_{\mathcal{X}^n} \exp\left\{ s\log\frac{dP_{X^n}}{dP_{\hat{X}^n}}(x^n)\right\} dP_{\hat{X}^n}(x^n).$$

Here, α_n and β_n are the type-I and type-II error probabilities, respectively.