

Learning a Bit of Information

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

October 12, 2023

The information processing task motivating the study of this lecture:

For a length- n sequence generated from one of two possible data sources, design a scheme to learn which one it comes from so that the probability of error vanishes the fastest to zero.

Fundamental Questions:

- How does the optimal performance scale with the number of *samples* n ?
- How to achieve that fundamental limit?

1 Basics of Binary Hypothesis Testing

2 Binary Hypothesis Testing in the Asymptotic Regime

3 Information Divergence

4 Error Exponents and Chernoff Information

Binary hypothesis testing: setup

- 1 Two hypotheses regarding the *observation* X , indexed by $\theta \in \{0, 1\}$:

$$\mathcal{H}_0 : X \sim P_0 \quad (\text{Null Hypothesis, } \theta = 0)$$

$$\mathcal{H}_1 : X \sim P_1 \quad (\text{Alternative Hypothesis, } \theta = 1)$$

- 2 Goal: design a decision rule ϕ to choose one of the two hypotheses based on the observed realization of X .

- 3 Two kinds of **probability of error**:

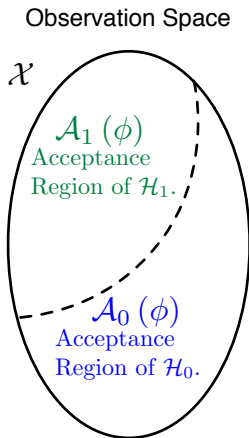
- Probability of **false alarm** (false positive; false rejection; type I error):

$$\pi_{1|0}(\phi) := \Pr\{\mathcal{H}_1 \text{ is chosen} \mid \mathcal{H}_0\}$$

- Probability of **miss detection** (false negative; false acceptance; type II error):

$$\pi_{0|1}(\phi) := \Pr\{\mathcal{H}_0 \text{ is chosen} \mid \mathcal{H}_1\}$$

A deterministic test \equiv two decision regions



A test $\phi : \mathcal{X} \rightarrow \{0, 1\}$ is equivalently characterized by its *acceptance (decision) regions*:

$$\mathcal{A}_{\hat{\theta}}(\phi) \equiv \phi^{-1}(\hat{\theta}) := \{x \in \mathcal{X} : \phi(x) = \hat{\theta}\}, \quad \hat{\theta} = 0, 1.$$

This gives an intuitive way to visualize the two types of probability of errors:

- Probability of **false alarm**:

$$\pi_{1|0}(\phi) = P_0(\mathcal{A}_1(\phi)) = 1 - P_0(\mathcal{A}_0(\phi))$$

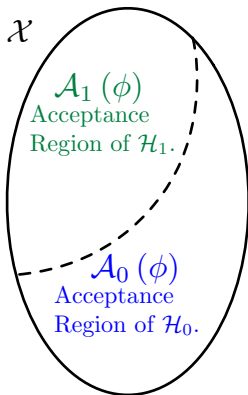
- Probability of **miss detection**:

$$\pi_{0|1}(\phi) = P_1(\mathcal{A}_0(\phi)) = 1 - P_1(\mathcal{A}_1(\phi))$$

In most cases there is a tradeoff between the two probability of errors.

Characterizing the trade-off

Observation Space



Finding the optimal trade-off can be formulated into an optimization problem:

$$\begin{aligned} & \underset{\mathcal{A}_0 \subseteq \mathcal{X}}{\text{minimize}} && P_1(\mathcal{A}_0) \\ & \text{subject to} && P_0(\mathcal{A}_0) \geq 1 - \epsilon \end{aligned}$$

- Similar to lossless source coding problem.
- Identical when $P_1 = \text{Unif}\{\mathcal{X}\}$ for $|\mathcal{X}| < \infty$.

The best set \mathcal{A}_0 can also be found by a greedy algorithm but not based on $P_0(x)$ anymore.

$P_0(x)/P_1(x)$ seems more reasonable as the score.

The greedy construction of the best acceptance region \mathcal{A}_0 based on the value of $P_0(x)/P_1(x)$ leads to a natural “likelihood ratio test” detailed next.

Likelihood ratio test

Definition 1 (Likelihood, Likelihood Ratio, Log Likelihood Ratio)

The *likelihood function* of θ given the observed data x is defined as¹ $L(\theta|x) := P_\theta(x)$. For binary hypothesis testing, it is useful to define the *likelihood ratio* and the *log likelihood ratio* (LLR) as follows:

$$\text{LR}(x) := \frac{L(1|x)}{L(0|x)} = \frac{P_1(x)}{P_0(x)} ; \quad \text{LLR}(x) := \log \text{LR}(x) = \log P_1(x) - \log P_0(x).$$

Definition 2 (Likelihood Ratio Test)

A likelihood ratio test (LRT) is a test ϕ_τ^{LRT} , parametrized by constants $\tau \in [0, +\infty]$ (called **threshold**), defined as $\phi_\tau^{\text{LRT}}(x) := \mathbb{1}\{\text{LR}(x) > \tau\}$, that is,

$$\mathcal{A}_0(\phi_\tau^{\text{LRT}}) = \{x \in \mathcal{X} \mid \tau P_0(x) \geq P_1(x)\}.$$

¹In this course, it is either a PMF $p_\theta(x)$ or a PDF $f_\theta(x)$.

Proposition 1

Consider the null hypothesis acceptance region of a test and that of an LRT, say, \mathcal{A} and \mathcal{A}^{LRT} respectively. Then, the two error probabilities cannot be simultaneously improved over the LRT, that is,

$$P_0(\mathcal{A}) \geq P_0(\mathcal{A}^{\text{LRT}}) \implies P_1(\mathcal{A}) \geq P_1(\mathcal{A}^{\text{LRT}}).$$

pf: Let the threshold of the LRT be τ : $\mathcal{A}^{\text{LRT}} = \{x \in \mathcal{X} \mid \tau P_0(x) \geq P_1(x)\}$. Then, by definition of a LRT,

$$\tau P_0(x) < P_1(x) \quad \forall x \in \mathcal{A} \setminus \mathcal{A}^{\text{LRT}}; \quad \tau P_0(x) \geq P_1(x) \quad \forall x \in \mathcal{A}^{\text{LRT}} \setminus \mathcal{A}.$$

Hence,

$$\begin{aligned} \tau (P_0(\mathcal{A}) - P_0(\mathcal{A}^{\text{LRT}})) &= \tau P_0(\mathcal{A} \setminus \mathcal{A}^{\text{LRT}}) - \tau P_0(\mathcal{A}^{\text{LRT}} \setminus \mathcal{A}) \\ &\leq P_1(\mathcal{A} \setminus \mathcal{A}^{\text{LRT}}) - P_1(\mathcal{A}^{\text{LRT}} \setminus \mathcal{A}) = P_1(\mathcal{A}) - P_1(\mathcal{A}^{\text{LRT}}). \end{aligned}$$

As a result, $P_1(\mathcal{A}) - P_1(\mathcal{A}^{\text{LRT}}) \geq \tau (P_0(\mathcal{A}) - P_0(\mathcal{A}^{\text{LRT}})) \geq 0$. □

Randomized tests

Randomized tests include deterministic tests as special cases.

Definition 3 (Randomized Test)

A randomized test declares $\hat{\theta} = 1$ with probability $\phi(x)$ and $\hat{\theta} = 0$ with probability $1 - \phi(x)$, where ϕ is a mapping $\phi : \mathcal{X} \rightarrow [0, 1]$.

Note: A randomized test is characterized by ϕ , as in deterministic tests.

Proposition 2

For a randomized test $\phi : \mathcal{X} \rightarrow [0, 1]$, the two types of probability of errors can be expressed as follows:

$$\begin{cases} \text{Probability of false alarm :} & \pi_{1|0}(\phi) = \mathbb{E}_{X \sim P_0} [\phi(X)] \\ \text{Probability of miss detection :} & \pi_{0|1}(\phi) = \mathbb{E}_{X \sim P_1} [1 - \phi(X)] \end{cases}$$

Randomized LRT

Definition 4 (Randomized Likelihood Ratio Test)

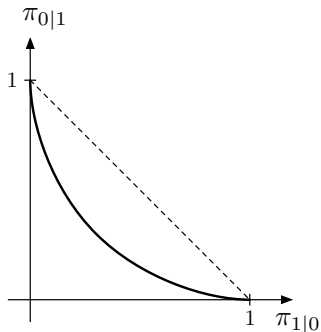
A (randomized) likelihood ratio test (LRT) is a test $\phi_{\tau,\gamma}$, parametrized by constants $\tau \in [0, +\infty]$ (called **threshold**) and $\gamma \in [0, 1]$, defined as follows:

$$\phi_{\tau,\gamma}(x) = \begin{cases} 1 & \text{if } \text{LR}(x) > \tau \\ \gamma & \text{if } \text{LR}(x) = \tau \\ 0 & \text{if } \text{LR}(x) < \tau \end{cases}$$

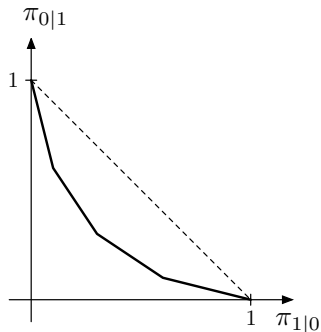
For a LRT $\phi_{\tau,\gamma}$, the probabilities of errors are

$$\begin{aligned} \pi_{1|0}(\phi_{\tau,\gamma}) &= P_0\{\text{LR}(X) > \tau\} + \gamma P_0\{\text{LR}(X) = \tau\} \\ \pi_{0|1}(\phi_{\tau,\gamma}) &= P_1\{\text{LR}(X) < \tau\} + (1 - \gamma) P_1\{\text{LR}(X) = \tau\} \end{aligned}$$

Varying τ and γ , there is a natural trade-off between $\pi_{1|0}$ and $\pi_{0|1}$.



Continuous Distribution



Discrete Distribution

What is the optimal trade-off curve?

What is the optimal test achieving it?

Optimal trade-off between error probabilities

To answer this question, we consider the following optimization problem:

NEYMAN-PEARSON PROBLEM

$$\begin{array}{ll} \underset{\phi: \mathcal{X} \rightarrow [0,1]}{\text{minimize}} & \pi_{0|1}(\phi) \\ \text{subject to} & \pi_{1|0}(\phi) \leq \epsilon \end{array}$$

(1)

Let $\varpi_{0|1}^*(\epsilon)$ denote the optimum value in (1).

Theorem 1 (Neyman-Pearson: (randomized) LRT is optimal)

For any $\epsilon \in [0, 1]$, $\varpi_{0|1}^(\epsilon)$ is attained by a (randomized) LRT ϕ_{τ^*, γ^*} , where the pair of parameters $(\tau^*, \gamma^*) \in [0, +\infty] \times [0, 1]$ is a solution to the following equation:*

$$P_0\{\text{LR}(X) > \tau\} + \gamma P_0\{\text{LR}(X) = \tau\} = \epsilon. \quad (2)$$

pf: First, it is not difficult to argue that (2) has a solution (τ^*, γ^*) for any $\epsilon \in [0, 1]$, due to the monotonicity of CDF of the likelihood ratio $\text{LR}(X)$.

Now, for any test ϕ , make the following observations:

- $L(1|x) - \tau^* L(0|x) > 0 \implies \phi_{\tau^*, \gamma^*}(x) = 1 \implies (\phi_{\tau^*, \gamma^*}(x) - \phi(x)) \geq 0.$
- $L(1|x) - \tau^* L(0|x) < 0 \implies \phi_{\tau^*, \gamma^*}(x) = 0 \implies (\phi_{\tau^*, \gamma^*}(x) - \phi(x)) \leq 0.$

Hence, $\forall x \in \mathcal{X}$, $(\phi_{\tau^*, \gamma^*}(x) - \phi(x)) (L(1|x) - \tau^* L(0|x)) \geq 0.$

Summing/integrating over all $x \in \mathcal{X}$, we get

$$(\pi_{0|1}(\phi) - \pi_{0|1}(\phi_{\tau^*, \gamma^*})) + \tau^* (\pi_{1|0}(\phi) - \pi_{1|0}(\phi_{\tau^*, \gamma^*})) \geq 0$$

Hence for any test ϕ with $\pi_{1|0}(\phi) \leq \epsilon = \pi_{1|0}(\phi_{\tau^*, \gamma^*})$, the probability of miss detection $\pi_{0|1}(\phi) \geq \pi_{0|1}(\phi_{\tau^*, \gamma^*})$. □

1 Basics of Binary Hypothesis Testing

2 Binary Hypothesis Testing in the Asymptotic Regime

3 Information Divergence

4 Error Exponents and Chernoff Information

i.i.d. Observations

So far: single shot observation from an arbitrary alphabet \mathcal{X} .

Next: multi-shot observation from a product space \mathcal{X}^n . Length- n sequence X^n drawn i.i.d. from one of the two distributions, and the two hypotheses are

$$\mathcal{H}_0 : X_i \stackrel{\text{i.i.d.}}{\sim} P_0, i = 1, 2, \dots, n \quad \equiv \quad X^n \sim P_0^{\otimes n}$$

$$\mathcal{H}_1 : X_i \stackrel{\text{i.i.d.}}{\sim} P_1, i = 1, 2, \dots, n \quad \equiv \quad X^n \sim P_1^{\otimes n}$$

The corresponding probability of errors for a test $\phi^{(n)}$ are denoted as

$$\begin{aligned}\pi_{1|0}^{(n)} &:= \Pr\{\mathcal{H}_1 \text{ is chosen} \mid \mathcal{H}_0\} = \mathbb{E}_{X^n \sim P_0^{\otimes n}} \left[\phi^{(n)}(X^n) \right] \\ \pi_{0|1}^{(n)} &:= \Pr\{\mathcal{H}_0 \text{ is chosen} \mid \mathcal{H}_1\} = \mathbb{E}_{X^n \sim P_1^{\otimes n}} \left[1 - \phi^{(n)}(X^n) \right]\end{aligned}$$

LRT under i.i.d. observations

With i.i.d. observations, the LLR of a sequence $x^n \in \mathcal{X}^n$ is

$$\text{LLR}(x^n) = \log \prod_{i=1}^n \frac{P_1(x_i)}{P_0(x_i)} = \sum_{i=1}^n \text{LLR}(x_i)$$

It turns out to be more convenient to consider the log likelihood ratio test $\varphi_{\eta, \gamma}$ (equivalent to LRT $\phi_{2^n, \gamma}$). Hence a (randomized) LRT becomes

$$\varphi_{\eta_n, \gamma_n}^{(n)}(x^n) = \begin{cases} 1/0 & \text{if } \sum_{i=1}^n \text{LLR}(x_i) \geq \eta_n \\ \gamma_n & \text{if } \sum_{i=1}^n \text{LLR}(x_i) = \eta_n \end{cases}.$$

Furthermore, the probabilities of type-I and type-II errors are

$$\pi_{1|0}^{(n)} = P_0^{\otimes n} \{ \sum_{i=1}^n \text{LLR}(X_i) > \eta_n \} + \gamma_n P_0^{\otimes n} \{ \sum_{i=1}^n \text{LLR}(X_i) = \eta_n \}$$

$$\pi_{0|1}^{(n)} = P_1^{\otimes n} \{ \sum_{i=1}^n \text{LLR}(X_i) \leq \eta_n \} - \gamma_n P_1^{\otimes n} \{ \sum_{i=1}^n \text{LLR}(X_i) = \eta_n \}$$

Asymptotic performance

Key: asymptotic behavior of the probability law of the sum of LLR's.

To highlight the role of each LLR, let $L_i \equiv \text{LLR}(X_i)$. L_1, \dots, L_n are also i.i.d., according to distribution $Q_\theta \equiv P_\theta \circ \text{LLR}^{-1}$.

For a LRT $\varphi^{(n)}$ with threshold η_n (on the LLR), the two probability of errors

$$\pi_{1|0}^{(n)} \approx Q_0^{\otimes n} \left\{ \sum_{i=1}^n L_i \geq \eta_n \right\}, \quad \pi_{0|1}^{(n)} \approx Q_1^{\otimes n} \left\{ \sum_{i=1}^n L_i \leq \eta_n \right\}.$$

Hence, by the WLLN, if the threshold η_n is chosen such that:

- $\eta_n = n(\mathbb{E}_{X \sim P_0}[\log \frac{P_1(X)}{P_0(X)}] + \delta)$ for some $\delta > 0$, and
- $\eta_n = n(\mathbb{E}_{X \sim P_1}[\log \frac{P_1(X)}{P_0(X)}] - \delta')$ for some $\delta' > 0$,

then both $\pi_{1|0}^{(n)}$ and $\pi_{0|1}^{(n)}$ vanish as $n \rightarrow \infty$.

Following the Neyman-Pearson criterion as in (2), if one only requires the type-I error probability $\pi_{1|0}^{(n)}$ to be less than or equal to any constant $\epsilon \in (0, 1)$, it suffices to choose $\eta_n \approx nE_{X \sim P_0}[\log \frac{P_1(X)}{P_0(X)}] =: -nD(P_0 \| P_1)$.

Moreover, suppose one is focused on the asymptotic behavior of the optimal type-II error probability

$$\varpi_{0|1}^*(n, \epsilon) := \min_{\phi: \mathcal{X}^n \rightarrow [0,1]} \pi_{0|1}^{(n)}(\phi) \quad \text{subject to } \pi_{1|0}^{(n)}(\phi) \leq \epsilon.$$

It turns out that $\varpi_{0|1}^*(n, \epsilon)$ vanishes exponentially fast, and the rate function

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \varpi_{0|1}^*(n, \epsilon) = D(P_0 \| P_1).$$

The above result looks similar to the lossless source coding problem, and hence can also be proved via a certain kind of *typicality*.

Stein's error exponent

Theorem 2 (Chernoff-Stein Lemma)

For any $\epsilon \in (0, 1)$, the error exponent of the type-II error probability is

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \varpi_{0|1}^*(n, \epsilon) = \mathbb{E}_{X \sim P_0} \left[\log \frac{P_0(X)}{P_1(X)} \right] =: D(P_0 \| P_1),$$

that is, $\varpi_{0|1}^*(n, \epsilon) = 2^{-n(D(P_0 \| P_1) + o(1))}$, or equivalently written as

$$\varpi_{0|1}^*(n, \epsilon) \doteq 2^{-nD(P_0 \| P_1)}.$$

To prove this theorem, we take a similar approach as in the lossless source coding by first defining a suitable typicality. Then for achievability, a deterministic test with acceptance region being a typical set is constructed.

For the converse part, we use typicality to control the “change of measure” from $P_1^{\otimes n}$ to $P_0^{\otimes n}$ and develop an exponentially tight lower bound.

Binary hypothesis testing

$$\begin{array}{ll} \underset{\mathcal{A} \subseteq \mathcal{X}^n}{\text{minimize}} & P_1^{\otimes n}(\mathcal{A}) \\ \text{subject to} & P_0^{\otimes n}(\mathcal{A}) \geq 1 - \epsilon \end{array}$$

Lossless source coding

$$\begin{array}{ll} \underset{\mathcal{A} \subseteq \mathcal{X}^n}{\text{minimize}} & |\mathcal{A}| \\ \text{subject to} & P^{\otimes n}(\mathcal{A}) \geq 1 - \epsilon \end{array}$$

Recall in the lossless source coding problem, on one hand, we leverage the WLLN to ensure that the constraint $P^{\otimes n}(\mathcal{A}) \geq 1 - \epsilon$, where

$$\mathcal{A} = \left\{ x^n \mid \frac{1}{n} \sum_{i=1}^n f(x_i) \approx \mathbb{E}_P[f(X)] \right\}.$$

On the other hand, to control $|\mathcal{A}|$ with $P^{\otimes n}(\mathcal{A}) \approx 1$, it suffices to control $P^{\otimes n}(x^n)$ for any sequence $x^n \in \mathcal{A}$.

Hence, we pick $f : x \mapsto \log P(x)$ so that $\forall x^n \in \mathcal{A}$,

$$P^{\otimes n}(x^n) \approx 2^{n\mathbb{E}_P[\log P(X)]}.$$

Consequently, $|\mathcal{A}| \approx 1/2^{n\mathbb{E}_P[\log P(X)]} = 2^{-n\mathbb{E}_P[\log P(X)]}$.

Binary hypothesis testing

$$\begin{aligned} & \underset{\mathcal{A} \subseteq \mathcal{X}^n}{\text{minimize}} && P_1^{\otimes n}(\mathcal{A}) \\ & \text{subject to} && P_0^{\otimes n}(\mathcal{A}) \geq 1 - \epsilon \end{aligned}$$

Lossless source coding

$$\begin{aligned} & \underset{\mathcal{A} \subseteq \mathcal{X}^n}{\text{minimize}} && |\mathcal{A}| \\ & \text{subject to} && P^{\otimes n}(\mathcal{A}) \geq 1 - \epsilon \end{aligned}$$

Now in the binary hypothesis testing problem, we again want to leverage the WLLN to ensure that the constraint $P_0^{\otimes n}(\mathcal{A}) \geq 1 - \epsilon$, where

$$\mathcal{A} = \left\{ x^n \mid \frac{1}{n} \sum_{i=1}^n f(x_i) \approx \mathbb{E}_{P_0}[f(X)] \right\}.$$

To control $P_1^{\otimes n}(\mathcal{A})$ with $P_0^{\otimes n}(\mathcal{A}) \approx 1$, now we should instead control $P_0^{\otimes n}(x^n)/P_1^{\otimes n}(x^n)$ for any sequence $x^n \in \mathcal{A}$.

Hence, we pick $f : x \mapsto \log(P_0(x)/P_1(x))$ so that $\forall x^n \in \mathcal{A}$,

$$P_0^{\otimes n}(x^n)/P_1^{\otimes n}(x^n) \approx 2^{n\mathbb{E}_{P_0}[\log \frac{P_0(X)}{P_1(X)}]}.$$

Consequently, $P_1^{\otimes n}(\mathcal{A}) \approx 1/2^{n\mathbb{E}_{P_0}[\log \frac{P_0(X)}{P_1(X)}]} = 2^{-n\mathbb{E}_{P_0}[\log \frac{P_0(X)}{P_1(X)}]}.$

Typicality

Definition 5

For two probability laws $P, Q \in \mathcal{P}(\mathcal{X})$ and $\delta > 0$, a δ -typical set of length- n sequences with respect to P, Q as

$$\mathcal{A}^{(n)}(P\|Q) := \left\{ x^n \in \mathcal{X}^n \mid \left| \frac{1}{n} \sum_{i=1}^n \log \frac{P(x_i)}{Q(x_i)} - D(P\|Q) \right| \leq \delta \right\}$$

Proposition 3

- 1 $\forall x^n \in \mathcal{A}_\delta^{(n)}(P\|Q), 2^{n(D(P\|Q)-\delta)} \leq P^{\otimes n}(x^n)/Q^{\otimes n}(x^n) \leq 2^{n(D(P\|Q)+\delta)}.$
- 2 $P^{\otimes n}(\mathcal{A}_\delta^{(n)}(P\|Q)) \geq 1 - \epsilon$ for n large enough.
- 3 $Q^{\otimes n}(\mathcal{A}_\delta^{(n)}(P\|Q)) \leq 2^{-n(D(P\|Q)-\delta)}.$
- 4 $Q^{\otimes n}(\mathcal{A}_\delta^{(n)}(P\|Q)) \geq (1 - \epsilon)2^{-n(D(P\|Q)+\delta)}$ for n large enough.

Achievability of Theorem 2

pf: Our goal is to show that there exist a sequence of tests $\{\phi^{(n)}\}$ with $\pi_{1|0}^{(n)} \leq \epsilon$ and $\forall \delta > 0, \pi_{0|1}^{(n)} \leq 2^{-n(D(P_0\|P_1)-\delta)}$ when n is large enough.

From the aforementioned typicality properties in Proposition 3, an obvious choice of the test is

$$\phi^{(n)}(x^n) = \begin{cases} 0, & \text{if } x^n \in \mathcal{A}_\delta^{(n)}(P_0\|P_1) \\ 1, & \text{otherwise} \end{cases}$$

By property 2 of Proposition 3, $\pi_{1|0}^{(n)}(\phi^{(n)}) \leq \epsilon$ when n is large enough.

Meanwhile, by property 3 of Proposition 3,

$$\pi_{0|1}^{(n)}(\phi^{(n)}) = P_1^{\otimes n} \left(\mathcal{A}_\delta^{(n)}(P_0\|P_1) \right) \leq 2^{-n(D(P_0\|P_1)-\delta)}.$$

This completes the proof of the achievability. □

Converse of Theorem 2

pf: For any given test $\phi^{(n)} : \mathcal{X}^n \rightarrow [0, 1]$ (could be randomized) with $\pi_{1|0}^{(n)}(\phi^{(n)}) \leq \epsilon$, let us first lower bound its type-II error probability as follows:

$$\begin{aligned}\pi_{0|1}^{(n)}(\phi^{(n)}) &= \mathbb{E}_{P_1^{\otimes n}} [1 - \phi^{(n)}(X^n)] \\ &\geq \mathbb{E}_{P_1^{\otimes n}} \left[\left(1 - \phi^{(n)}(X^n)\right) \mathbb{1} \left\{ X^n \in \mathcal{A}_{\delta'}^{(n)}(P_0 \| P_1) \right\} \right], \end{aligned} \quad (3)$$

where the last step is taking the expectation over the typical set $\mathcal{A}_{\delta'}^{(n)}(P_0 \| P_1)$.

Next, to leverage the fact that $\pi_{1|0}^{(n)}(\phi^{(n)}) = \mathbb{E}_{P_0^{\otimes n}} [\phi^{(n)}(X^n)] \leq \epsilon$ to further lower bound (3), we would like to change the distribution in the expectation from $P_1^{\otimes n}$ to $P_0^{\otimes n}$. That is, for each $x^n \in \mathcal{A}_{\delta'}^{(n)}(P_0 \| P_1)$ in the expectation, we change $P_1^{\otimes n}(x^n)$ to $P_0^{\otimes n}(x^n)$. By property 1 of Proposition 3, change of measure leads to a further lower bound:

$$(3) \geq \mathbb{E}_{P_0^{\otimes n}} \left[\left(1 - \phi^{(n)}(X^n)\right) \mathbb{1} \left\{ X^n \in \mathcal{A}_{\delta'}^{(n)}(P_0 \| P_1) \right\} \right] 2^{-n(D(P_0 \| P_1) + \delta')}. \quad (4)$$

Further observe that

$$\begin{aligned} & \mathbb{E}_{P_0^{\otimes n}} \left[\left(1 - \phi^{(n)}(X^n) \right) \mathbb{1} \left\{ X^n \in \mathcal{A}_{\delta'}^{(n)}(P_0 \| P_1) \right\} \right] \\ &= P_0^{\otimes n} \left(\mathcal{A}_{\delta'}^{(n)}(P_0 \| P_1) \right) - \mathbb{E}_{P_0^{\otimes n}} \left[\phi^{(n)}(X^n) \mathbb{1} \left\{ X^n \in \mathcal{A}_{\delta'}^{(n)}(P_0 \| P_1) \right\} \right] \\ &\geq (1 - \epsilon') - \mathbb{E}_{P_0^{\otimes n}} \left[\phi^{(n)}(X^n) \right] \quad \text{for } n \text{ large enough.} \end{aligned} \tag{5}$$

Here the last step is due to property 2 of Proposition 3.

Combining (3) – (5), and the fact that $\pi_{1|0}^{(n)}(\phi^{(n)}) = \mathbb{E}_{P_0^{\otimes n}} [\phi^{(n)}(X^n)] \leq \epsilon$, we arrive at

$$\pi_{0|1}^{(n)}(\phi^{(n)}) \geq (1 - \epsilon' - \epsilon) 2^{-n(D(P_0 \| P_1) + \delta')} \quad \text{for } n \text{ large enough.}$$

The proof is complete by choosing ϵ', δ' properly. □

Reflections

- The key to the proof is the **weak law of large numbers (WLLN)**, in both achievability and converse. WLLN suffices because we only impose a *constant* requirement on one probability of error. It is expected to have exponential decay of the other one, due to change of measure.
- $D(P_0 \| P_1)$ characterizes the error exponent of mistaking \mathcal{H}_1 as \mathcal{H}_0 . It can be understood as telling how dissimilar P_0 is from P_1 , in the problem of learning a piece of binary information.
 $D(P_0 \| P_1)$ is called the **information divergence of P_0 from P_1** .
- Note that WLLN no longer suffices when both error probabilities are imposed with *exponential decaying* requirements.
The speed of concentration, in addition to the fact of concentration itself, is needed. This is what **large deviation principle** will tell us.

Summary: binary hypothesis testing

- For binary hypothesis testing (the task of determining from data X which distribution out of two possible ones (P_0 and P_1) it follows), the likelihood ratio $LR(x) = \frac{P_1(x)}{P_0(x)}$ plays a central role, and it is a sufficient statistic.
- For binary hypothesis testing with i.i.d. sampled data, the optimal error exponent is characterized by the **information divergence** of one distribution from the other:

$$\varpi_{0|1}^*(n, \epsilon) \doteq 2^{-nD(P_0 \| P_1)}.$$

Next: definition, interpretation, and properties of information divergence

- 1 Basics of Binary Hypothesis Testing
- 2 Binary Hypothesis Testing in the Asymptotic Regime
- 3 Information Divergence**
- 4 Error Exponents and Chernoff Information

Information divergence

Recap: Consider the task of determining from i.i.d. sampled data X_1, \dots, X_n which law out of two possible ones (P_0 and P_1) they follow.

Given the constraint that the probability of mistaking P_0 for P_1 is $\leq \epsilon$, the probability of mistaking P_1 for P_0

$$\varpi_{0|1}^*(n, \epsilon) \doteq 2^{-nD(P_0\|P_1)}.$$

So, getting one more sample, roughly speaking, the probability of error in the logarithmic scale decrease by an amount $D(P_0\|P_1)$. This decrease in the log scale is the **information divergence of P_0 from P_1** .

Hence, information divergence of P_0 from P_1 can be viewed as a measure of the level of “dissimilarity” of P_0 from P_1 .

The general definition of information divergence involves the Radon-Nikodym derivative in measure theory, and hence omitted here. Instead, we focus on two important cases and give the equivalent definitions accordingly.

Information Divergence (Kullback–Leibler Divergence, Relative Entropy)

- \mathcal{A} is discrete: in this case, P and Q are PMF's, and

$$D(P\|Q) := \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)} = \sum_{a \in \text{supp} P} P(a) \log \frac{P(a)}{Q(a)}$$

with the convention $0 \log \frac{0}{q} = 0 \forall q \in [0, 1]$, and $p \log \frac{p}{0} = \infty \forall p \in (0, 1]$.

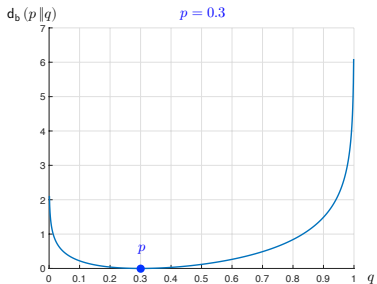
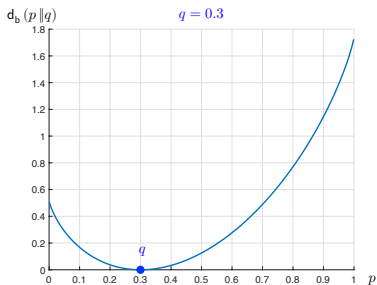
Note: if $\text{supp} P \setminus \text{supp} Q \neq \emptyset$, then $D(P\|Q) = +\infty$.

- $\mathcal{A} = \mathbb{R}^d$ and P and Q are PDF's: if $\text{supp} P \setminus \text{supp} Q$ has zero volume (Lebesgue measure = 0), then

$$D(P\|Q) := \int_{\mathbb{R}^d} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{Q(\mathbf{x})} d\mathbf{x} = \int_{\text{supp} P} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{Q(\mathbf{x})} d\mathbf{x}.$$

with the convention $0 \log \frac{0}{q} = 0 \forall q \in [0, 1]$, and $p \log \frac{p}{0} = \infty \forall p \in (0, 1]$.

If $\text{supp} P \setminus \text{supp} Q$ has non-zero volume, $D(P\|Q) = +\infty$.



Example 1 (Binary divergence function)

Let $\mathbf{p} = (1 - p, p)$ and $\mathbf{q} = (1 - q, q)$, $p, q \in (0, 1)$. The information divergence of \mathbf{p} from \mathbf{q} is called the **binary divergence function** $d_b(p \parallel q)$, where

$$\begin{aligned} d_b(p \parallel q) &:= D(\mathbf{p} \parallel \mathbf{q}) \\ &= p \log \frac{p}{q} + (1 - p) \log \frac{(1-p)}{(1-q)}. \end{aligned}$$

Exercise 1

Prove the following statements.

- 1 $d_b(p \parallel \frac{1}{2}) = 1 - H_b(p)$.
- 2 $d_b(p \parallel q)$ is convex in (p, q) .
- 3 $d_b(p \parallel q) \geq 2(p - q)^2 \log_2 e$.

Non-negativity of information divergence

Theorem 3

$D(P\|Q) \geq 0$, with equality iff $P(x) = Q(x)$ for all $x \in \mathcal{X}$.

pf: For simplicity, we only prove the discrete case:

$$\begin{aligned} -D(P\|Q) &= E_{X \sim P} \left[\log \frac{Q(X)}{P(X)} \right] \\ &\stackrel{\text{Jensen}}{\leq} \log E_{X \sim P} \left[\frac{Q(X)}{P(X)} \right] = \log \sum_{a \in \text{supp}P} Q(a) \leq \log 1 = 0, \end{aligned}$$

where the first inequality is due to the fact that $x \mapsto \log x$ is strictly concave.

The second inequality becomes equality iff $\text{supp}Q \subseteq \text{supp}P$. The first one becomes equality iff $\forall x \in \text{supp}P$, $P(x) = c Q(x)$ for some constant c .

As a result, $D(P\|Q) = 0$ iff $\text{supp}Q = \text{supp}P$ and $\forall x \in \text{supp}P$, $P(x) = Q(x)$. □

Data processing

We always process data to extract information.

$$X \sim P_X \longrightarrow \boxed{W_{Y|X}} \longrightarrow Y \sim P_Y$$

$$X \sim Q_X \longrightarrow \boxed{W_{Y|X}} \longrightarrow Y \sim Q_Y$$

In general, the processing of input data X into output data Y can be viewed as a *stochastic mapping*, equivalently characterized by a *conditional probability law* $W_{Y|X}(y|x)$ for $y \in \mathcal{Y}$ and $x \in \mathcal{X}$.

The following theorem, known as the **data processing inequality**, tells us the the dissimilarity after the processing cannot be increased.

Theorem 4 (Data processing cannot increase divergence)

$$D(P_X \| Q_X) \geq D(P_Y \| Q_Y).$$

pf: Let the joint laws of (X, Y) in the upper branch and the lower branch be $P_{X,Y}$ and $Q_{X,Y}$ respectively.

Nota that $\forall x, y, \frac{P_{X,Y}(x,y)}{Q_{X,Y}(x,y)} = \frac{P_X(x)W_{Y|X}(y|x)}{Q_X(x)W_{Y|X}(y|x)} = \frac{P_X(x)}{Q_X(x)}$. Hence,

$$\begin{aligned} -D(P_X \| Q_X) &= -D(P_{X,Y} \| Q_{X,Y}) \\ &= E_{Y \sim P_Y} \left[E_{X \sim P_{X|Y}} \left[\log \frac{Q_{X,Y}(X,Y)}{P_{X,Y}(X,Y)} \middle| Y \right] \right] \\ &\leq E_{Y \sim P_Y} \left[\log E_{X \sim P_{X|Y}} \left[\frac{Q_{X,Y}(X,Y)}{P_{X,Y}(X,Y)} \middle| Y \right] \right] \\ &= E_{Y \sim P_Y} \left[\log \frac{Q_Y(Y)}{P_Y(Y)} \right] = -D(P_Y \| Q_Y), \end{aligned}$$

where the inequality is due to Jensen. □

Theorem 4 is also called the **data processing inequality (DPI)**.

Conditional information divergence

Similar to entropy, one can define conditional relative entropy (information divergence), which measures the dissimilarity of one conditional law from another conditional law, averaged over a marginal law.

Definition 6 (Conditional Information Divergence)

Let $P_{Y|X}$ and $Q_{Y|X}$ be two conditional laws of Y given X and P_X be a marginal law of X . The conditional information divergence of $P_{Y|X}$ from $Q_{Y|X}$ conditioned on P_X is defined as

$$D(P_{Y|X} \| Q_{Y|X} | P_X) := E_{X \sim P_X} [D(P_{Y|X}(\cdot|X) \| Q_{Y|X}(\cdot|X))] .$$

Exercise 2 (Nonnegativity of Conditional Divergence)

Show $D(P_{Y|X} \| Q_{Y|X} | P_X) \geq 0$.

Chain rule for information divergence

Information divergence, as a measure of dissimilarity, is additive.

Theorem 5 (Chain Rule for Information Divergence)

Let $P_{X,Y}$ and $Q_{X,Y}$ be two joint laws of (X, Y) with conditional laws of Y given X being $P_{Y|X}$ and $Q_{Y|X}$ respectively and marginal laws of X being P_X and Q_X respectively. Then,

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_{Y|X} \| Q_{Y|X} | P_X) + D(P_X \| Q_X).$$

pf: By definition and the telescoping factorization of joint probability law. □

Exercise 3

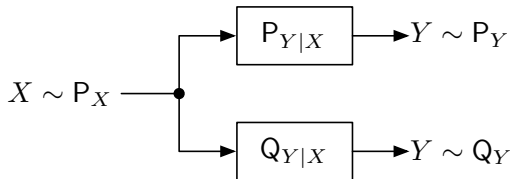
Let P_{X^n} and Q_{X^n} be joint laws of X^n with conditional laws and marginal laws defined similarly as above. Show that

$$D(P_{X^n} \| Q_{X^n}) = \sum_{i=1}^n D(P_{X_i | X^{i-1}} \| Q_{X_i | X^{i-1}} | P_{X^{i-1}}).$$

Furthermore, suppose Q_{X^n} is a product law: $Q_{X^n}(x^n) = \prod_{i=1}^n Q_{X_i}(x_i)$. Show that

$$D(P_{X^n} \| Q_{X^n}) = D\left(P_{X^n} \left\| \prod_{i=1}^n P_{X_i}\right.\right) + \sum_{i=1}^n D(P_{X_i} \| Q_{X_i}).$$

Conditioning increases information divergence



Theorem 6 (Conditioning Increases Information Divergence)

Let $P_{Y|X}$ and $Q_{Y|X}$ be two conditional laws of Y given X and P_X be a marginal law of X . Let P_Y and Q_Y denote the marginal laws of Y after passing $X \sim P_X$ through $P_{Y|X}$ and $Q_{Y|X}$ respectively. Then,

$$D(P_Y \| Q_Y) \leq D(P_{Y|X} \| Q_{Y|X} | P_X), \text{ with equality iff } D(P_{X|Y} \| Q_{X|Y} | P_Y) = 0.$$

Conditioned on the common “input” law, dissimilarity between $P_{Y|X}$ and $Q_{Y|X}$ is **higher** than that between the “output” laws P_Y and Q_Y .

pf: Let the joint law of (X, Y) in the upper and lower branches be $P_{X,Y}$ and $Q_{X,Y}$ respectively.

Note that the marginal laws of X under the two joint laws are both P_X .

By the Chain Rule (Theorem 5),

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_{Y|X} \| Q_{Y|X} | P_X) + D(P_X \| P_X) = D(P_{Y|X} \| Q_{Y|X} | P_X)$$

$$D(P_{X,Y} \| Q_{X,Y}) = D(P_{X|Y} \| Q_{X|Y} | P_Y) + D(P_Y \| Q_Y)$$

The proof is complete by comparing the above two and the non-negativity of $D(P_{X|Y} \| Q_{X|Y} | P_Y)$. □

Convexity of information divergence

Theorem 7

$D(P\|Q)$ is convex in (P, Q) , that is, for any $\lambda \in [0, 1]$ and P_0, P_1, Q_0, Q_1 ,

$$D(P_\lambda\|Q_\lambda) \leq (1 - \lambda)D(P_0\|Q_0) + \lambda D(P_1\|Q_1),$$

where $P_\lambda = (1 - \lambda)P_0 + \lambda P_1$ and $Q_\lambda = (1 - \lambda)Q_0 + \lambda Q_1$.

The proof is immediate: this is a special case of “Conditioning Increases Information Divergence” (Theorem 6) a binary input law.

Variational representation of information divergence

Information divergence is a measure of dissimilarity from one law to another. It has wide applications in statistics and machine learning.

The following variational representation of information divergence due to Donsker and Varadhan is quite useful for approximating information divergence using optimization methods.

Theorem 8 (Donsker-Varadhan)

For two probability laws P and Q on a common alphabet \mathcal{X} , if $D(P\|Q) < \infty$, then for any $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $E_{X \sim Q}[2^{f(X)}] < \infty$, $E_{X \sim P}[f(X)]$ exists. Furthermore,

$$D(P\|Q) = \max_{f: \mathcal{X} \rightarrow \mathbb{R}} E_{X \sim P}[f(X)] - \log E_{X \sim Q}[2^{f(X)}] \quad \text{s.t. } E_{X \sim Q}[2^{f(X)}] < \infty.$$

pf: There are two directions to prove: “ \leq ” and “ \geq ”.

For the direction “ \leq ”, it suffices to find $f : \mathcal{X} \rightarrow \mathbb{R}$ with $E_{X \sim Q}[2^{f(X)}] < \infty$ such that $D(P\|Q) = E_{X \sim P}[f(X)] - \log E_{X \sim Q}[2^{f(X)}]$.

An obvious choice is $f : x \mapsto \log \frac{P(x)}{Q(x)}$ (note that $E_{X \sim Q} \left[\frac{P(X)}{Q(X)} \right] = 1$).

For the direction “ \geq ”, need to show for any $f : \mathcal{X} \rightarrow \mathbb{R}$ with $E_{X \sim Q}[2^{f(X)}] < \infty$, $D(P\|Q) \geq E_{X \sim P}[f(X)] - \log E_{X \sim Q}[2^{f(X)}]$, that is,

$$E_{X \sim P} \left[\log P(X) - \underbrace{\left(f(X) + \log Q(X) - \log E_{X \sim Q}[2^{f(X)}] \right)}_{(*)} \right] \geq 0.$$

A deeper look into the term $(*)$ reveals that $(*) = \log Q_f(X)$ for another law Q_f , defined as $Q_f(x) := \frac{2^{f(x)} Q(x)}{E_{X \sim Q}[2^{f(X)}]}$ (tilted law of Q).

The proof is complete by $D(P\|Q_f) \geq 0$. □

Summary: information divergence

- Information divergence (or KL divergence in short) is a measure of dissimilarity of one law from another.
- The measure of information in discriminating two hypotheses.
- Nonnegative and convex, asymmetric, not satisfying triangle inequality.
- Variational representation of information divergence (Donsker & Varadhan).
- Data processing inequality, increase by conditioning, and chain rule.

- 1 Basics of Binary Hypothesis Testing
- 2 Binary Hypothesis Testing in the Asymptotic Regime
- 3 Information Divergence
- 4 Error Exponents and Chernoff Information

Trade-off between $\pi_{1|0}^{(n)}$ and $\pi_{0|1}^{(n)}$ as $n \rightarrow \infty$

Recall: Chernoff-Stein Lemma (Theorem 2) tells us when we only impose a **constant** requirement on one of the two error probabilities, the other vanishes exponentially fast:

$$\pi_{1|0}^{(n)} \leq \epsilon, \pi_{0|1}^{(n)} \doteq 2^{-nD(P_0 \| P_1)} \quad (6)$$

$$\pi_{0|1}^{(n)} \leq \epsilon, \pi_{1|0}^{(n)} \doteq 2^{-nD(P_1 \| P_0)} \quad (7)$$

It turns out that both error probabilities can vanish exponentially fast as the number of samples $n \rightarrow \infty$.

To formalize, let us begin with the following definitions.

Definition 7

A pair of type-I and type-II exponents (e_1, e_2) is *achievable* if \exists a sequence of tests $\{\phi^{(n)} \mid n \in \mathbb{N}\}$ such that the corresponding $(\pi_{1|0}^{(n)}, \pi_{0|1}^{(n)})$ satisfy

$$\liminf_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log \pi_{1|0}^{(n)} \right\} \geq e_1, \quad \liminf_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log \pi_{0|1}^{(n)} \right\} \geq e_2.$$

The optimal type-II error exponent subject to a type-I exponent e_1 is

$$e_2^*(e_1) := \sup \{e_2 \mid (e_1, e_2) \text{ is achievable}\}.$$

Chernoff-Stein Lemma suggests that

$$e_2^*(e_1) = \begin{cases} D(P_0 \| P_1), & \text{if } e_1 = 0 \\ 0, & \text{if } e_1 = D(P_1 \| P_0) \end{cases}.$$

It remains to characterize the optimal type-II error exponent function for $e_1 \in (0, D(P_1 \| P_0))$.

There are various ways to find the optimal trade-off between the exponents.

One approach: LRT is optimal, so it suffices to characterize its asymptotic performance using **large deviation theory**.

Main tool: **Cramér's Theorem**, which says that the probability the empirical average of n i.i.d. random variables deviating from its mean by a small constant vanishes exponentially fast with n , and the rate function (exponent) is the **Legendre transform of the cumulant generating function of the distribution**, evaluated at this small constant.

By applying Cramér's Theorem to the sum of LLR's, the optimal trade-off between exponents can be quantified *parametrically*. See Chapter 15 and 16 of Polyanskiy and Wu, or Chapter 3.2 of B. C. Levy's book².

Pros: its generality (discrete and continuous distributions).

Cons: requires some knowledge about convex analysis, Fenchel's duality theorem, and tilted distribution.

²B. C. Levy, Principles of Signal Detection and Parameter Estimation, Springer Verlag, 2008.

We do not take this approach. Instead, we employ the **Method of Types** (the name coined by I. Csiszár³), a set of tools that are quite powerful in the context of discrete probabilities.

This approach can be traced back to W. Hoeffding⁴, the first to quantify the optimal trade-off between exponents for binary hypothesis testing.

To me, it is more insightful than the Cramér's-Theorem approach. It clearly tells why the optimal trade-off between error exponents can be expressed as the result of an **information projection** problem (variational representation).

Starting from there, we then solve the information projection problem (a convex program) by observing that its optimizer must be a **tilted distribution**.

Next, we give an equivalent parametric characterization (due to R. E. Blahut⁵) which is useful in finding the asymptotic performance in the Bayesian and Minimax settings.

³I. Csiszár, "The Method of Types," IEEE Transactions on Information Theory, vol. 44, no. 6, 1998.

⁴W. Hoeffding, "Asymptotically Optimal Tests for Multinomial Distributions," Ann. Math. Statist., vol. 36, pp. 1916-1921, 1956.

⁵R. E. Blahut, "Hypothesis Testing and Information Theory," IEEE Transactions on Information Theory, vol. 20, no. 4, 1974.

Type

Definition 8 (Type/Empirical Distribution)

For a sequence $x^n \in \mathcal{X}^n$ where $\mathcal{X} = \{a_1, a_2, \dots, a_d\}$, the number of occurrence of a symbol $a \in \mathcal{X}$ in the sequence is defined as

$$n(a|x^n) := \sum_{i=1}^n \mathbb{1}\{x_i = a\}, \quad a \in \mathcal{X}.$$

The *type* (empirical distribution/histogram) of the sequence is a d -dimensional probability vector defined as the collection of all d frequencies of occurrence:

$$\hat{\mathbf{p}}(x^n) := \frac{1}{n}(n(a_1|x^n), n(a_2|x^n), \dots, n(a_d|x^n)).$$

It belongs to the probability simplex \mathcal{P}_d . Hence, when the context is about probability laws, we interchangeably use $\hat{\mathbf{P}}_{x^n}$ to represent its type.

Note: The type of a sequence is a deterministic function of the sequence. Nothing random here.

Note that the possible types of sequences of length $n \in \mathbb{N}$ just form a strict subset of the probability simplex, and the number of them is polynomial in n .

Definition 9 (Collection of Types)

For a given length $n \in \mathbb{N}$, we use $\hat{\mathcal{P}}_{n,d}$ to denote the collection of all possible “ n -types” of length- n sequences:

$$\hat{\mathcal{P}}_{n,d} := \left\{ \left(\frac{i_1}{n}, \frac{i_2}{n}, \dots, \frac{i_d}{n} \right) \mid \forall i_1, \dots, i_d \in \mathbb{N} \cup \{0\}, i_1 + \dots + i_d = n \right\}.$$

Exercise 4

Show that

$$|\hat{\mathcal{P}}_{n,d}| = \binom{n+d-1}{d-1} \leq (n+1)^d.$$

Type class

For a type $\hat{P} \in \hat{\mathcal{P}}_{n,d}$, there are many length- n sequences with the same type \hat{P} . The collection of such sequences is called a *type class*.

Definition 10 (Type Class)

For a type $\hat{P} \in \hat{\mathcal{P}}_{n,d}$, the type class of \hat{P} , $\mathcal{T}_n(\hat{P})$, is the collection of length- n sequences with type \hat{P} :

$$\mathcal{T}_n(\hat{P}) := \left\{ x^n \in \mathcal{X}^n \mid \hat{P}_{x^n} = \hat{P} \right\}.$$

It is important to count the number of sequences with the same type. The size of the type class of $\mathcal{T}_n(\hat{p})$ can be characterized explicitly as the following combinatorial expression which gives little insight:

$$|\mathcal{T}_n(\hat{p})| = \frac{n!}{(n\hat{p}_1)!(n\hat{p}_2)!\cdots(n\hat{p}_d)!}, \text{ where } \hat{p} = (\hat{p}_1, \dots, \hat{p}_d).$$

It turns out that $|\mathcal{T}_n(\hat{P})|$ is roughly $2^{nH(\hat{P})}$, to within a poly- n factor.

Probability of a memoryless sequence

Type is a natural way to classify sequences when we are concerned about the probability of an i.i.d. generated sequence.

Proposition 4 (Probability of an i.i.d. sequence depends only on its type)

For an i.i.d. generated sequence X^n with $X_i \stackrel{i.i.d.}{\sim} P$, $i = 1, 2, \dots, n$,

$$P^{\otimes n}\{X^n = x^n\} = 2^{-n(H(\hat{P}_{x^n}) + D(\hat{P}_{x^n} \| P))}.$$

pf: For notational convenience, let us denote $P\{X = a\}$ as $P(a)$. We prove this proposition by direct calculation:

$$\begin{aligned} P^{\otimes n}(x^n) &= \prod_{i=1}^n P(x_i) = \prod_{l=1}^d P(a_l)^{n(a_l|x^n)} = 2^{-n \sum_{l=1}^d \frac{n(a_l|x^n)}{n} \log \frac{1}{P(a_l)}} \\ &= 2^{-n \sum_{l=1}^d \left\{ \hat{P}_{x^n}(a_l) \log \frac{1}{\hat{P}_{x^n}(a_l)} + \hat{P}_{x^n}(a_l) \log \frac{\hat{P}_{x^n}(a_l)}{P(a_l)} \right\}} \\ &= 2^{-n[H(\hat{P}_{x^n}) + D(\hat{P}_{x^n} \| P)]}. \end{aligned}$$



Size of a type class

The size of a type class can be approximately characterized by its entropy, which is quite similar to the set of typical sequences.

Proposition 5 (Bounds on the Size of a Type Class)

For a type $\hat{P} \in \hat{\mathcal{P}}_{n,d}$, $\frac{1}{|\hat{\mathcal{P}}_{n,d}|} 2^{nH(\hat{P})} \leq |\mathcal{T}_n(\hat{P})| \leq 2^{nH(\hat{P})}$.

pf: Our proof is based on a probabilistic argument. Construct an i.i.d. sequence $X^n \sim \hat{P}^{\otimes n}$. By Proposition 4, we have

$$\hat{P}^{\otimes n}(x^n) = 2^{-n[H(\hat{P}_{x^n}) + D(\hat{P}_{x^n} \parallel \hat{P})]}.$$

If $x^n \in \mathcal{T}_n(\hat{P})$, then $\hat{P}_{x^n} = \hat{P}$ and $\hat{P}^{\otimes n}(x^n) = 2^{-nH(\hat{P})}$. Hence,

$$1 \geq \sum_{x^n \in \mathcal{T}_n(\hat{P})} \hat{P}^{\otimes n}(x^n) = |\mathcal{T}_n(\hat{P})| 2^{-nH(\hat{P})} \implies |\mathcal{T}_n(\hat{P})| \leq 2^{nH(\hat{P})},$$

and the upper bound is proved.

To prove the lower bound, observe that for any type $\hat{Q} \in \hat{\mathcal{P}}_{n,d}$, the probability that X^n (following $\hat{P}^{\otimes n}$) falls into this type class $\mathcal{T}_n(\hat{Q})$ is maximized when $\hat{Q} = \hat{P}$:

$$\hat{P}^{\otimes n}(\mathcal{T}_n(\hat{P})) \geq \hat{P}^{\otimes n}(\mathcal{T}_n(\hat{Q})), \quad \forall \hat{Q} \in \hat{\mathcal{P}}_{n,d}. \quad (8)$$

The proof of this is left as exercise (Imre Csiszár: Simple algebra.)

By (8) we have

$$\begin{aligned} 1 &= \sum_{\hat{Q} \in \hat{\mathcal{P}}_{n,d}} \hat{P}^{\otimes n}(\mathcal{T}_n(\hat{Q})) \leq |\hat{\mathcal{P}}_{n,d}| \hat{P}^{\otimes n}(\mathcal{T}_n(\hat{P})) \\ &= |\hat{\mathcal{P}}_{n,d}| |\mathcal{T}_n(\hat{P})| 2^{-nH(\hat{P})} \implies |\mathcal{T}_n(\hat{P})| \geq \frac{1}{|\hat{\mathcal{P}}_{n,d}|} 2^{nH(\hat{P})}. \end{aligned} \quad \square$$

Remark: Since $|\hat{\mathcal{P}}_{n,d}|$ polynomial in n , the size of the typical set $|\mathcal{T}_n(\hat{P})|$ has the same exponential order as $2^{nH(\hat{P})}$, that is, $|\mathcal{T}_n(\hat{P})| \doteq 2^{nH(\hat{P})}$.

Exercise 5

Prove (8) by direct calculation and note that $m!/n! \geq n^{m-n}$ for $m, n \in \mathbb{N}$.

Probability of a type class

Proposition 4 tells us the probability of a sequence of a given type under i.i.d. distribution. Proposition 5 gives exponentially tight bounds for the size of a type class. As a result, we have exponentially tight bounds for the probability of a type class under i.i.d. distribution as follows.

Proposition 6 (Probability of a type class and information divergence)

For an i.i.d. generated sequence X^n with $X_i \stackrel{i.i.d.}{\sim} P$, $i = 1, 2, \dots, n$,

$$\frac{1}{|\hat{\mathcal{P}}_{n,d}|} 2^{-nD(\hat{P}||P)} \leq P^{\otimes n} \left\{ X^n \in \mathcal{T}_n(\hat{P}) \right\} \leq 2^{-nD(\hat{P}||P)}.$$

In other words, $P^{\otimes n}(\mathcal{T}_n(\hat{P})) \doteq 2^{-nD(\hat{P}||P)}$.

The proof is straightforward and hence omitted.

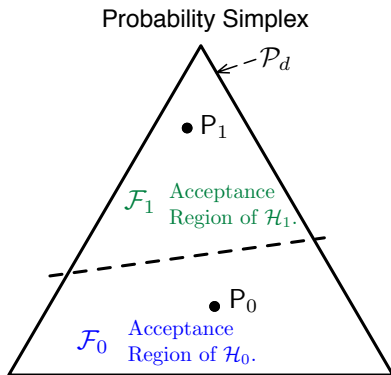
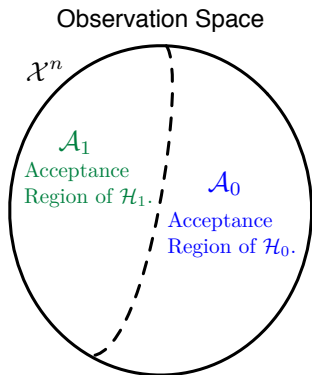
LRT \equiv Comparing divergences of the type

Interestingly, LRT for testing \mathcal{H}_0 vs. \mathcal{H}_1 from i.i.d. $X^n \sim P_\theta^{\otimes n}$ can be cast as comparing the information divergences of the type \hat{P}_{X^n} from P_0 and P_1 :

$$\begin{aligned}\text{LLR}(x^n) &= \sum_{i=1}^n \log P_1(x_i) - \log P_0(x_i) \\ &= \sum_{a \in \mathcal{X}} (\log P_1(a) - \log P_0(a)) n(a|x^n) \\ &= n \sum_{a \in \mathcal{X}} (\log P_1(a) - \log P_0(a)) \hat{P}_{x^n}(a) \\ &= n \left\{ \sum_{a \in \mathcal{X}} \left(\log \frac{\hat{P}_{x^n}(a)}{P_0(a)} \right) \hat{P}_{x^n}(a) - \sum_{a \in \mathcal{X}} \left(\log \frac{\hat{P}_{x^n}(a)}{P_1(a)} \right) \hat{P}_{x^n}(a) \right\} \\ &= n \left\{ D(\hat{P}_{x^n} \| P_0) - D(\hat{P}_{x^n} \| P_1) \right\}.\end{aligned}$$

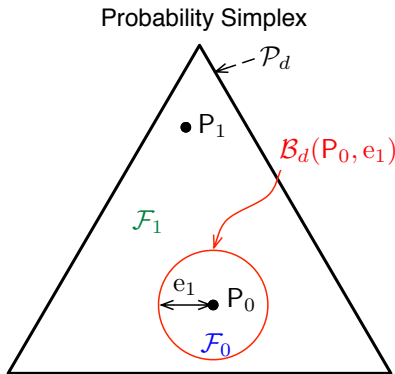
Hence, LRT is equivalent to comparing $D(\hat{P}_{x^n} \| P_1)$ and $D(\hat{P}_{x^n} \| P_0)$:

- If the type of the sequence x^n , \hat{P}_{x^n} , is “closer” to P_1 : decide $\hat{\theta} = 1$.
- Otherwise, decide $\hat{\theta} = 0$.



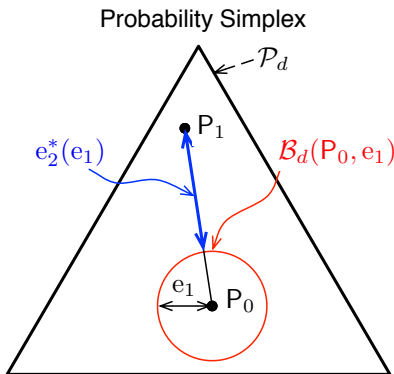
$$\{x^n \in \mathcal{A}_i \iff \text{decide } \mathcal{H}_i\} \iff \{\hat{\mathbf{P}}_{x^n} \in \mathcal{F}_i \iff \text{decide } \mathcal{H}_i\}.$$

It turns out this intuition gives a simple test (by W. Hoeffding, 1965) that is asymptotically optimal.



Information divergence “ball” $\mathcal{B}_d(P_0, e_1) := \{P \in \mathcal{P}_d \mid D(P \| P_0) \leq e_1\}$.

Hoeffding’s test: $\phi^{(n)}(x^n) = \mathbb{1} \left\{ \hat{P}_{x^n} \notin \mathcal{B}_d(P_0, e_1) \right\}$.



Information divergence “ball” $\mathcal{B}_d(P_0, e_1) := \{P \in \mathcal{P}_d \mid D(P \| P_0) \leq e_1\}$.

Hoeffding’s test is asymptotically optimal and achieves the optimal type-II error exponent subject to a type-I exponent e_1 (see the next theorem).

Characterization of the optimal error exponent

Theorem 9

For any $e_1 \in (0, D(P_1 \| P_0))$, $\boxed{e_2^*(e_1) = \min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1)}$. In words, it is the “divergence” of a “divergence ball” with radius e_1 centered at P_0 from P_1 .

pf: Proof Outline:

- **Achievability:** construct tests $\{\phi^{(n)}\}$ with $\liminf_{n \rightarrow \infty} \{-\frac{1}{n} \log \pi_{1|0}^{(n)}\} \geq e_1$ and

$$\liminf_{n \rightarrow \infty} \{-\frac{1}{n} \log \pi_{0|1}^{(n)}\} \geq \min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1)$$

- **Converse:** for any tests $\{\phi^{(n)}\}$ with $\liminf_{n \rightarrow \infty} \{-\frac{1}{n} \log \pi_{1|0}^{(n)}\} \geq e_1$, show that

$$\limsup_{n \rightarrow \infty} \{-\frac{1}{n} \log \pi_{0|1}^{(n)}\} \leq \min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1)$$

Achievability: Hoeffding's test (type-based): $\phi^{(n)}(x^n) = \mathbb{1} \left\{ \hat{P}_{x^n} \notin \mathcal{B}_d(P_0, e_1) \right\}$.

Type-I error probability: we need $\pi_{1|0}^{(n)} \leq 2^{-ne_1}$. Using method of types,

$$\begin{aligned} \pi_{1|0}^{(n)} &= \sum_{\hat{P} \in \hat{\mathcal{P}}_{n,d}: D(\hat{P} \| P_0) > e_1} P_0^{\otimes n}(\mathcal{T}_n(\hat{P})) \\ &\leq \sum_{\hat{P} \in \hat{\mathcal{P}}_{n,d}: D(\hat{P} \| P_0) > e_1} 2^{-nD(\hat{P} \| P_0)} \\ &< |\hat{\mathcal{P}}_{n,d}| 2^{-ne_1} \doteq 2^{-ne_1}. \end{aligned}$$

Type-II error probability: we need $\pi_{0|1}^{(n)} \leq 2^{-n \min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1)}$. Using method of types, similarly we get

$$\begin{aligned} \pi_{0|1}^{(n)} &\leq \sum_{\hat{P} \in \hat{\mathcal{P}}_{n,d} \cap \mathcal{B}_d(P_0, e_1)} 2^{-nD(\hat{P} \| P_1)} \\ &\doteq 2^{-n \min_{\hat{P} \in \hat{\mathcal{P}}_{n,d} \cap \mathcal{B}_d(P_0, e_1)} D(\hat{P} \| P_1)} \\ &\leq 2^{-n \min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1)}. \end{aligned}$$

Converse: Here we are given a sequence of tests (possibly randomized tests) $\{\phi^{(n)}\}$ with $\pi_{1|0}^{(n)} \leq 2^{-ne_1}$, and our goal is to show that

$$\pi_{0|1}^{(n)} \geq 2^{-n(\min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1))}.$$

Let us first lower bound the type-II error probability: for any type $\hat{P} \in \hat{\mathcal{P}}_{n,d}$,

$$\begin{aligned} \pi_{0|1}^{(n)}(\phi^{(n)}) &= E_{P_1^{\otimes n}} [1 - \phi^{(n)}(X^n)] \\ &\geq E_{P_1^{\otimes n}} \left[\left(1 - \phi^{(n)}(X^n)\right) \mathbb{1} \left\{ X^n \in \mathcal{T}_n(\hat{P}) \right\} \right], \end{aligned} \quad (9)$$

where the last step is taking the expectation over the type class $\mathcal{T}_n(\hat{P})$.

Next, to leverage the fact that $\pi_{1|0}^{(n)}(\phi^{(n)}) = E_{P_0^{\otimes n}} [\phi^{(n)}(X^n)] \leq 2^{-ne_1}$ to further lower bound (9), we would like to do change-the-measure from $P_1^{\otimes n}$ to $P_0^{\otimes n}$.

This is easy for sequences of the same type: by Proposition 4, $\forall x^n \in \mathcal{T}_n(\hat{P})$,

$$P_1^{\otimes n}(x^n)/P_0^{\otimes n}(x^n) = 2^{n(D(\hat{P} \| P_0) - D(\hat{P} \| P_1))}.$$

With the change-of-measure within the type class,

$$(9) = \mathbb{E}_{P_0^{\otimes n}} \left[\left(1 - \phi^{(n)}(X^n) \right) \mathbb{1} \left\{ X^n \in \mathcal{T}_n(\hat{P}) \right\} \right] 2^{n(D(\hat{P} \| P_0) - D(\hat{P} \| P_1))} \quad (10)$$

Further observe that

$$\begin{aligned} & \mathbb{E}_{P_0^{\otimes n}} \left[\left(1 - \phi^{(n)}(X^n) \right) \mathbb{1} \left\{ X^n \in \mathcal{T}_n(\hat{P}) \right\} \right] \\ &= P_0^{\otimes n}(\mathcal{T}_n(\hat{P})) - \mathbb{E}_{P_0^{\otimes n}} \left[\phi^{(n)}(X^n) \mathbb{1} \left\{ X^n \in \mathcal{T}_n(\hat{P}) \right\} \right] \\ &\geq 2^{-nD(\hat{P} \| P_0)} - 2^{-ne_1}. \end{aligned} \quad (11)$$

The last step is due to Proposition 6 and the assumption that $\pi_{1|0}^{(n)} \leq 2^{-ne_1}$.

Combining (9) – (11), we arrive at

$$\pi_{0|1}^{(n)}(\phi^{(n)}) \geq \left(1 - 2^{-n(e_1 - D(\hat{P} \| P_0))} \right) 2^{-nD(\hat{P} \| P_1)}.$$

The above bound works for any n -type \hat{P} . We shall pick \hat{P} in the interior of $\mathcal{B}_d(P_0, e_1)$ so that the pre-factor converges to 1.

It remains to show that for $\epsilon > 0$, can pick a $\hat{P} \in \mathcal{B}_d(P_0, e_1 - \delta) \cap \hat{\mathcal{P}}_{n,d}$ such that

$$D(\hat{P} \| P_1) \leq \min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1) + \epsilon$$

for a nicely chosen $\delta > 0$ and sufficiently large n . This is possible due to the following reasons:

- $D(P \| P_1)$ is convex and continuous in P . Moreover, the constraint set $\mathcal{B}_d(P_0, e_1)$ is a closed convex set. Hence, it is possible to choose a sufficiently small $\delta > 0$ such that

$$\min_{P \in \mathcal{B}_d(P_0, e_1 - \delta)} D(P \| P_1) \leq \min_{P \in \mathcal{B}_d(P_0, e_1)} D(P \| P_1) + \epsilon/2.$$

- Again, since $D(P \| P_1)$ is continuous in P , we can choose a sufficiently large n such that there exists $\hat{P} \in \mathcal{B}_d(P_0, e_1 - \delta) \cap \hat{\mathcal{P}}_{n,d}$ with

$$D(\hat{P} \| P_1) \leq \min_{P \in \mathcal{B}_d(P_0, e_1 - \delta)} D(P \| P_1) + \epsilon/2.$$

The proof is now complete. □

Information projection

Theorem 9 tells us that the optimal tradeoff can be expressed as a minimization problem which is reminiscent of **projection onto a “ball”**:

$$\begin{aligned} \min_{\mathbf{p} \in \mathbb{R}^d} \quad & \underbrace{\sum_{i=1}^d p_i \log \frac{p_i}{p_{1,i}}}_{D(\mathbf{P} \parallel \mathbf{P}_1)} \\ \text{subject to} \quad & \underbrace{\sum_{i=1}^d p_i \log \frac{p_i}{p_{0,i}}}_{D(\mathbf{P} \parallel \mathbf{P}_0)} \leq e_1 \\ & \sum_{i=1}^d p_i = 1, \quad p_i \geq 0, \quad \forall i = 1, \dots, d \end{aligned} \tag{12}$$

with the understanding that $\mathbf{p} \leftrightarrow \mathbf{P}$, $\mathbf{p}_0 \leftrightarrow \mathbf{P}_0$, and $\mathbf{p}_1 \leftrightarrow \mathbf{P}_1$.

Since this is a convex program, one can write down the KKT conditions and find the optimal solution. The minimizer turns out to take the following form:

$$p_i^* = \frac{p_{0,i}^{1-\lambda} p_{1,i}^\lambda}{\sum_{i=1}^d p_{0,i}^{1-\lambda} p_{1,i}^\lambda}, \quad i = 1, \dots, d, \quad \text{for some } \lambda \in [0, 1].$$

The parametrized distribution P_λ found by solving the KKT conditions,

$$P_\lambda(a) := \frac{P_0(a)^{1-\lambda} P_1(a)^\lambda}{\sum_{a \in \mathcal{X}} P_0(a)^{1-\lambda} P_1(a)^\lambda}, \quad \forall a \in \mathcal{X},$$

is actually a *tilted distribution* of P_0 by recognizing

$$P_\lambda(a) \propto P_0(a)^{1-\lambda} P_1(a)^\lambda = 2^{\lambda \log \frac{P_1(a)}{P_0(a)}} P_0(a).$$

This implies that $\log \frac{P_\lambda(a)}{P_0(a)} = \lambda \log \frac{P_1(a)}{P_0(a)} - \log \mathbb{E}_{X \sim P_0} \left[2^{\lambda \log \frac{P_1(X)}{P_0(X)}} \right]$.

Hence, when λ increases from 0 to 1, P_λ naturally **interpolates from P_0 to P_1 with increasing divergence from P_0** (proof is left as exercise), unless $P_0 = P_1$. Consequently, one can find a unique $\lambda^* \in (0, 1)$ so that

$$D(P_{\lambda^*} \| P_0) = e_1. \quad (13)$$

Exercise 6

When $P_0 \neq P_1$, show that $D(P_\lambda \| P_0)$ is a continuous and strictly increasing function of λ , for $\lambda \in [0, 1)$. Hence, the choice in (13) is unique.

Instead of solving the KKT condition which is quite mechanical and gives little insight, let us directly prove the optimal solution.

Theorem 10

Let λ^* be chosen as in (13). For any $P \in \mathcal{B}_d(P_0, e_1)$,

$$D(P \| P_1) - D(P_{\lambda^*} \| P_1) \geq \frac{1}{\lambda^*} D(P \| P_{\lambda^*}) \geq 0.$$

Hence, $e_2^*(e_1) = D(P_{\lambda^*} \| P_1)$. In other words, the optimal trade-off between the two exponents e_1 and e_2 is characterized by the parametric curve below:

$$\left\{ (D(P_\lambda \| P_0), D(P_\lambda \| P_1)) \mid \lambda \in (0, 1) \right\}. \quad (14)$$

pf: The following identity will be used twice in the rest of this proof:

$$\begin{aligned} D(P \| P_\lambda) &= E_P [\log P(X) - (1 - \lambda)P_0(X) - \lambda P_1(X)] + \log E_{P_0} \left[2^{\lambda \log \frac{P_1(X)}{P_0(X)}} \right] \\ &= (1 - \lambda)D(P \| P_0) + \lambda D(P \| P_1) + \log E_{P_0} \left[2^{\lambda \log \frac{P_1(X)}{P_0(X)}} \right]. \end{aligned} \quad (15)$$

From (15), we can plug in $P \leftarrow P_\lambda$ and get

$$(1 - \lambda)D(P_\lambda \| P_0) + \lambda D(P_\lambda \| P_1) = -\log E_{P_0} \left[2^{\lambda \log \frac{P_1(X)}{P_0(X)}} \right]. \quad (16)$$

From (15) and (16),

$$\lambda D(P \| P_1) = D(P \| P_\lambda) - (1 - \lambda)D(P \| P_0) + (1 - \lambda)D(P_\lambda \| P_0) + \lambda D(P_\lambda \| P_1).$$

In particular, for any $P \in \mathcal{B}_d(P_0, e_1)$, since $D(P \| P_0) \leq e_1$, we have

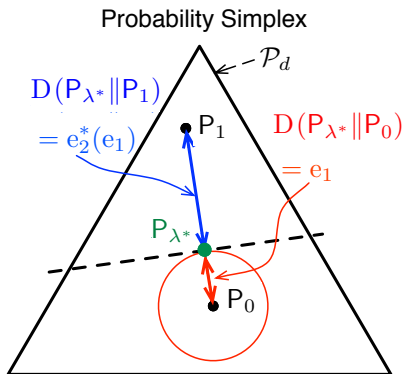
$$\begin{aligned} \lambda D(P \| P_1) &\geq D(P \| P_\lambda) - (1 - \lambda)e_1 + (1 - \lambda)D(P_\lambda \| P_0) + \lambda D(P_\lambda \| P_1) \\ &= D(P \| P_\lambda) - (1 - \lambda)D(P_{\lambda^*} \| P_0) + (1 - \lambda)D(P_\lambda \| P_0) + \lambda D(P_\lambda \| P_1), \end{aligned}$$

where the last equality is due to the choice in (13).

The above inequality holds for all $\lambda \in [0, 1)$. Plugging in $\lambda \leftarrow \lambda^*$, we get

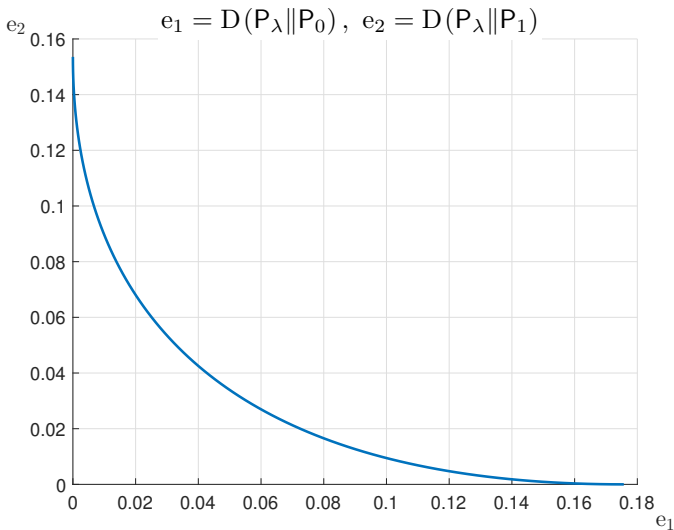
$$\lambda^* D(P \| P_1) \geq D(P \| P_{\lambda^*}) + \lambda^* D(P_{\lambda^*} \| P_1),$$

which completes the proof. □



The tilted distribution P_λ naturally interpolates from P_0 to P_1 with increasing divergence from P_0 , as λ grows from 0 towards 1.

By symmetry, you can also start from P_1 to tilt the distribution towards P_0 .



The two corner points are $(D(P_1 \| P_0), 0)$ and $(0, D(P_0 \| P_1))$ respectively.

Chernoff Information

With the characterization of the optimal trade-off between the two error exponents, $e_2^*(e_1)$, we are able to tackle the Bayes error probability and the minimax error probability.

For the Bayes setting, let the prior distribution be $\pi \equiv (\pi_0, \pi_1)$. The Bayes error probability for the binary hypothesis testing problem with n i.i.d. samples is

$$P_e^*(\pi, n) := \min_{\phi: \mathcal{X}^n \rightarrow [0,1]} \left\{ \pi_0 \pi_{1|0}^{(n)}(\phi) + \pi_1 \pi_{0|1}^{(n)}(\phi) \right\}. \quad (17)$$

Meanwhile, for the minimax setting, the minimax error probability is

$$\bar{P}_e^*(n) := \min_{\phi: \mathcal{X}^n \rightarrow [0,1]} \left\{ \max \left\{ \pi_{1|0}^{(n)}(\phi), \pi_{0|1}^{(n)}(\phi) \right\} \right\}. \quad (18)$$

Since both $\pi_{1|0}^{(n)} \doteq 2^{-ne_1}$ and $\pi_{0|1}^{(n)} \doteq 2^{-ne_2}$ vanishes exponentially with n , both the Bayes error probability (17) and the minimax one (18) vanish exponentially with n too, and the exponent is just $\min\{e_1, e_2\}$.

As a result, the optimal error exponents for the Bayes error probability (17) and the minimax one (18) are both equal to

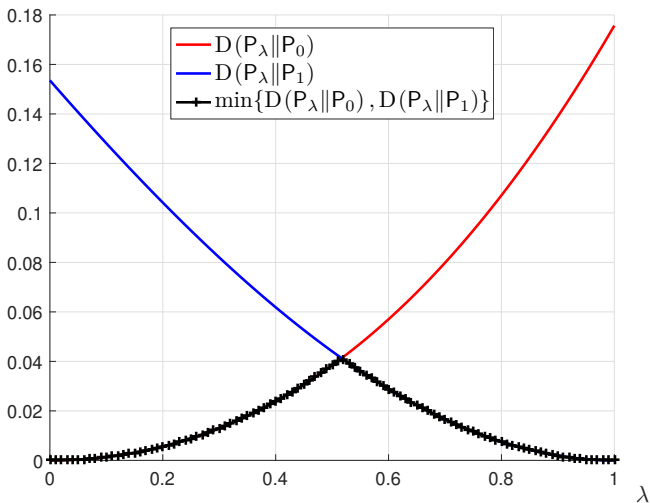
$$\max_{e_1 \in (0, D(P_1 \| P_0))} \min\{e_1, e_2^*(e_1)\} = \max_{\lambda \in (0, 1)} \min\{D(P_\lambda \| P_0), D(P_\lambda \| P_1)\}. \quad (19)$$

The last equality is due to Theorem 10.

By Exercise 6, it is readily seen that $D(P_\lambda \| P_0)$ and $D(P_\lambda \| P_1)$ are strictly increasing and decreasing functions of λ respectively. Hence, the maximizing λ^* in the maximization problem in (19) occurs at the unique point of intersection of the two curves $D(P_\lambda \| P_0)$ and $D(P_\lambda \| P_1)$, that is,

$$D(P_{\lambda^*} \| P_0) = D(P_{\lambda^*} \| P_1),$$

as long as the above equality holds for some $\lambda^* \in (0, 1)$, which is obviously true (see picture in the next slide).



The two curves only intersect at one point $\lambda^* \in (0, 1)$.

Now, with the identity in (16), we see that the optimizing λ^* satisfies

$$D(P_{\lambda^*} \| P_0) = D(P_{\lambda^*} \| P_1) = -\log \left(\sum_{a \in \mathcal{X}} P_0(a)^{1-\lambda^*} P_1(a)^{\lambda^*} \right).$$

Plugging back to the maximization problem (19), we see that the optimal error exponents for the Bayes error probability (17) and the minimax error probability (18) are both *upper bounded* by

$$CI(P_0, P_1) := \max_{\lambda \in (0,1)} \underbrace{-\log \sum_{a \in \mathcal{X}} P_0(a)^{1-\lambda} P_1(a)^{\lambda}}_{f(\lambda)}. \quad (20)$$

Furthermore, it can be shown that $\lambda^* = \arg \max_{\lambda \in (0,1)} f(\lambda)$. This is because

$$f'(\lambda) = D(P_{\lambda} \| P_1) - D(P_{\lambda} \| P_0), \quad (\text{exercise})$$

leading to $f''(\lambda) < 0$ for $\lambda \in (0, 1)$, and hence λ^* satisfying $f'(\lambda^*) = 0$ is the unique maximizer of (20).

The following theorem summarizes the above discussion.

Theorem 11

For binary hypothesis testing with n i.i.d. samples from P_0 versus P_1 , the optimal error exponent for the Bayes and the minimax error probability is

$$\lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log P_e^*(\pi, n) \right\} = \lim_{n \rightarrow \infty} \left\{ -\frac{1}{n} \log \bar{P}_e^*(n) \right\} = \text{CI}(P_0, P_1),$$

the Chernoff Information between P_0 and P_1 , as defined in (20).

From binary to m -ary hypothesis testing

Beyond binary hypothesis testing problems, one can consider m -ary hypothesis testing problems, where $\theta \in \Theta = \{1, 2, \dots, m\}$. It is possible to define achievable error exponents and characterize the region of all achievable tuple of error exponents. See the work by E. Tuncel⁶.

In some applications, there is an additional “rejection” option, where you can reject all m hypotheses, and the trade-off is between the error probability and the rejection probability. The characterization of optimal trade-off between exponents can be found the work by N. Grigoryan *et al.*⁷.

For us, in most situations it suffices to get a good upper bound on the average error probability. We will introduce some of these bounds in the context of noisy channel coding in the next unit.

⁶E. Tuncel, “On Error Exponents in Hypothesis Testing,” IEEE Transactions on Information Theory, vol. 51, no. 8, 2005.

⁷N. Grigoryan, A. N. Harutyunyan, S. Voloshynovskiy, and O. J. Kova “On Multiple Hypothesis Testing with Rejection Option,” Proceedings of the IEEE Information Theory Workshop, 2011.

Summary: error exponents

- Trade-off between error exponents in binary hypothesis testing can be quantified by information projection: $e_2^*(e_1) = \min_{P: D(P\|P_0) \leq e_1} D(P\|P_1)$.
- Equivalently, the region of achievable exponents are upper bounded by a parametric curve $\{(D(P_\lambda\|P_0), D(P_\lambda\|P_1)) \mid \lambda \in (0, 1)\}$, where P_λ denotes the tilted distribution

$$P_\lambda(a) := \frac{P_0(a)^{1-\lambda} P_1(a)^\lambda}{\sum_{a \in \mathcal{X}} P_0(a)^{1-\lambda} P_1(a)^\lambda}, \quad \forall a \in \mathcal{X}.$$

- Development based on large deviation theory or method of types.
- Optimal exponent of the Bayes error probability and the minimax error probability is the Chernoff information

$$CI(P_0, P_1) := \max_{\lambda \in (0, 1)} \left\{ \log \frac{1}{\sum_{a \in \mathcal{X}} P_0(a)^{1-\lambda} P_1(a)^\lambda} \right\}.$$