Learning a Bit of Information

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

September 28, 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?

Binary hypothesis testing: setup

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

$$\mathcal{H}_0 : X \sim \mathsf{P}_0 \quad \text{(Null Hypothesis, } \theta = 0\text{)}$$
$$\mathcal{H}_1 : X \sim \mathsf{P}_1 \quad \text{(Alternative Hypothesis, } \theta = 1\text{)}$$

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

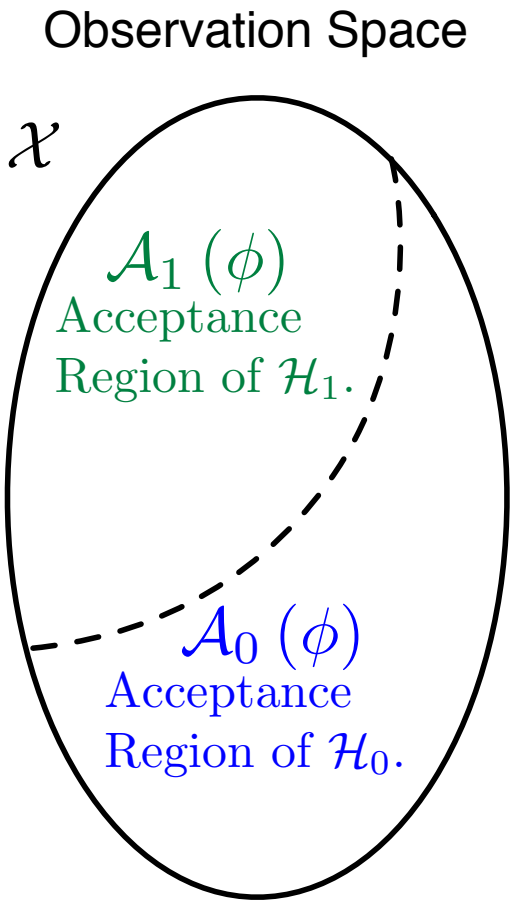
3. Two kinds of probability of errors:
   - Probability of false alarm (false positive; false rejection; type I error):

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

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

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

A deterministic test $\equiv$ two decision regions



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

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

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

- Probability of false alarm:

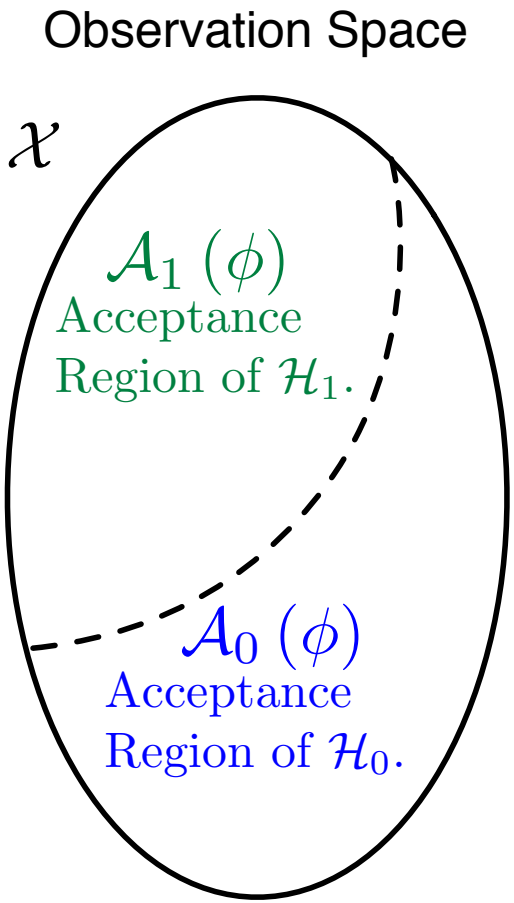
$$\pi_{1|0}(\phi) = \mathsf{P}_0(\mathcal{A}_1(\phi)) = 1 - \mathsf{P}_0(\mathcal{A}_0(\phi))$$

- Probability of miss detection:

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

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

Characterizing the trade-off



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

$$\begin{aligned} \underset{\mathcal{A}_0 \subseteq \mathcal{X}}{\text{minimize}} \quad & \mathsf{P}_1(\mathcal{A}_0) \\ \text{subject to} \quad & \mathsf{P}_0(\mathcal{A}_0) \geq 1-\epsilon \end{aligned}$$

- Similar to the loss source coding problem.
- Identical when $\mathsf{P}_1 = \mathrm{Unif}\{\mathcal{X}\}$ for $|\mathcal{X}| < \infty$.

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

$\mathsf{P}_0(x)/\mathsf{P}_1(x)$ seems more reasonable as the score.

The greedy construction of the best acceptance region $\mathcal{A}_0$ based on the value of $\mathsf{P}_0(x)/\mathsf{P}_1(x)$ 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 as[1] $\mathsf{L}(\theta|x) := \mathsf{P}_\theta(x)$. For binary hypothesis testing, it is useful to define the *likelihood ratio* and the *log likelihood ratio* (LLR) as follows:

$$\mathsf{LR}(x) := \frac{\mathsf{L}(1|x)}{\mathsf{L}(0|x)} = \frac{\mathsf{P}_1(x)}{\mathsf{P}_0(x)} \; ; \quad \mathsf{LLR}(x) := \log \mathsf{LR}(x) = \log \mathsf{P}_1(x) - \log \mathsf{P}_0(x).$$

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 $\phi_\tau^{\mathsf{LRT}}(x) := \mathbb{1}\{\mathsf{LR}(x) > \tau\}$, that is,

$$\mathcal{A}_0(\phi_\tau^{\mathsf{LRT}}) = \left\{x \in \mathcal{X} \,\middle|\, \tau \mathsf{P}_0(x) \geq \mathsf{P}_1(x)\right\}.$$

[1] In this course, it is either a PMF $\mathsf{p}_\theta(x)$ or a PDF $\mathsf{f}_\theta(x)$.

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,

$$\mathsf{P}_0(\mathcal{A}) \geq \mathsf{P}_0(\mathcal{A}^{LRT}) \implies \mathsf{P}_1(\mathcal{A}) \geq \mathsf{P}_1(\mathcal{A}^{LRT}).$$

pf: Let the threshold of the LRT be τ: $\mathcal{A}^{\mathsf{LRT}} = \left\{x \in \mathcal{X} \,\middle|\, \tau \mathsf{P}_0(x) \geq \mathsf{P}_1(x)\right\}$.
Then, by definition of a LRT,

$$\tau \mathsf{P}_0(x) < \mathsf{P}_1(x) \quad \forall\, x \in \mathcal{A} \setminus \mathcal{A}^{\mathsf{LRT}}\,; \quad \tau \mathsf{P}_0(x) \geq \mathsf{P}_1(x) \quad \forall\, x \in \mathcal{A}^{\mathsf{LRT}} \setminus \mathcal{A}.$$

Hence,

$$\begin{aligned}\tau\left(\mathsf{P}_0(\mathcal{A}) - \mathsf{P}_0(\mathcal{A}^{\mathsf{LRT}})\right) &= \tau \mathsf{P}_0\left(\mathcal{A} \setminus \mathcal{A}^{\mathsf{LRT}}\right) - \tau \mathsf{P}_0\left(\mathcal{A}^{\mathsf{LRT}} \setminus \mathcal{A}\right) \\ &\leq \mathsf{P}_1\left(\mathcal{A} \setminus \mathcal{A}^{\mathsf{LRT}}\right) - \mathsf{P}_1\left(\mathcal{A}^{\mathsf{LRT}} \setminus \mathcal{A}\right) = \mathsf{P}_1(\mathcal{A}) - \mathsf{P}_1(\mathcal{A}^{\mathsf{LRT}}).\end{aligned}$$

As a result, $\mathsf{P}_1(\mathcal{A}) - \mathsf{P}_1(\mathcal{A}^{\mathsf{LRT}}) \geq \tau\left(\mathsf{P}_0(\mathcal{A}) - \mathsf{P}_0(\mathcal{A}^{\mathsf{LRT}})\right) \geq 0$. □

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 $\phi : \mathcal{X} \to [0, 1]$.

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

Proposition 2

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

$$\begin{cases} \textit{Probability of false alarm}: & \pi_{1|0}(\phi) = \mathsf{E}_{X \sim \mathsf{P}_0}\left[\phi(X)\right] \\ \textit{Probability of miss detection}: & \pi_{0|1}(\phi) = \mathsf{E}_{X \sim \mathsf{P}_1}\left[1 - \phi(X)\right] \end{cases}$$

Randomized LRT

Definition 4 (Randomized Likelihood Ratio Test)

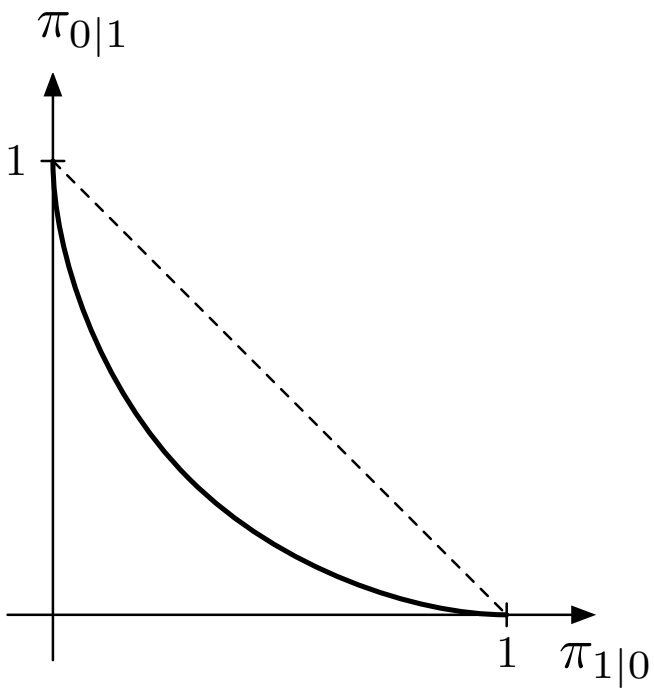
A (randomized) likelihood ratio test (LRT) is a test $\phi_{\tau,\gamma}$, parametrized by constants $\tau \in [0, +\infty]$ (called threshold) and $\gamma \in [0, 1]$, defined as follows:

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

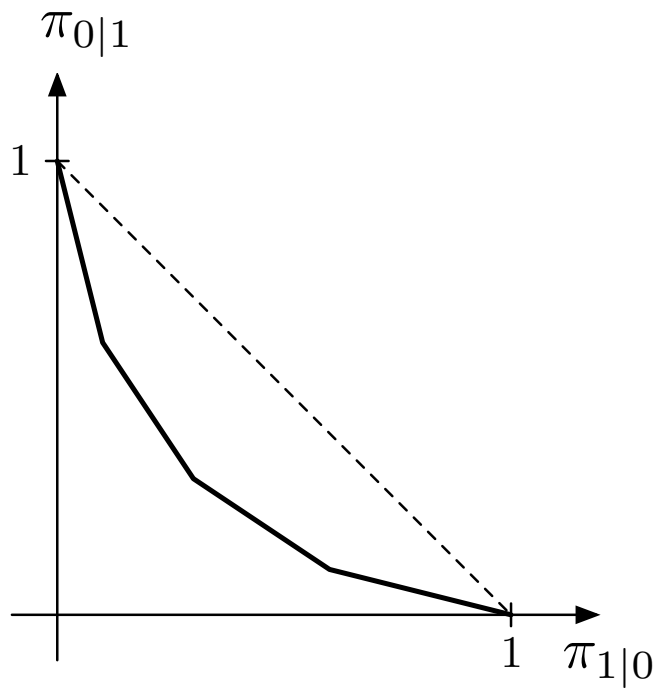
For a LRT $\phi_{\tau,\gamma}$, the probabilities of errors are

$$\pi_{1|0}(\phi_{\tau,\gamma}) = \mathsf{P}_0\{\mathsf{LR}(X) > \tau\} + \gamma\, \mathsf{P}_0\{\mathsf{LR}(X) = \tau\}$$
$$\pi_{0|1}(\phi_{\tau,\gamma}) = \mathsf{P}_1\{\mathsf{LR}(X) < \tau\} + (1-\gamma)\, \mathsf{P}_1\{\mathsf{LR}(X) = \tau\}$$

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:

$$\underline{\text{Neyman-Pearson Problem}}$$

$$\begin{aligned} &\underset{\phi:\mathcal{X}\to[0,1]}{\text{minimize}} && \pi_{0|1}(\phi) \\ &\text{subject to} && \pi_{1|0}(\phi) \le \epsilon \end{aligned} \tag{1}$$

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:

$$\mathsf{P}_0\{\mathsf{LR}(X) > \tau\} + \gamma\, \mathsf{P}_0\{\mathsf{LR}(X) = \tau\} = \epsilon. \tag{2}$$

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 $\mathsf{LR}(X)$.

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

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

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

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

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

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

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

$$\begin{aligned} \mathcal{H}_0 : X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_0,\ i = 1, 2, \ldots, n \quad &\equiv \quad X^n \sim \mathsf{P}_0^{\otimes n} \\ \mathcal{H}_1 : X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_1,\ i = 1, 2, \ldots, n \quad &\equiv \quad X^n \sim \mathsf{P}_1^{\otimes n} \end{aligned}$$

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

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

LRT under i.i.d. observations

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

$$\mathrm{LLR}\left(x^n\right) = \log \prod_{i=1}^{n} \frac{\mathsf{P}_1(x_i)}{\mathsf{P}_0(x_i)} = \sum_{i=1}^{n} \mathrm{LLR}(x_i)$$

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

$$\varphi^{(n)}_{\eta_n,\gamma_n}\left(x^n\right) = \begin{cases} 1/0 & \text{if } \sum_{i=1}^{n} \mathrm{LLR}\left(x_i\right) \gtrless \eta_n \\ \gamma_n & \text{if } \sum_{i=1}^{n} \mathrm{LLR}\left(x_i\right) = \eta_n \end{cases}.$$

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

$$\pi^{(n)}_{1|0} = \mathsf{P}_0^{\otimes n}\left\{\sum_{i=1}^{n} \mathrm{LLR}(X_i) > \eta_n\right\} + \gamma_n \mathsf{P}_0^{\otimes n}\left\{\sum_{i=1}^{n} \mathrm{LLR}(X_i) = \eta_n\right\}$$
$$\pi^{(n)}_{0|1} = \mathsf{P}_1^{\otimes n}\left\{\sum_{i=1}^{n} \mathrm{LLR}(X_i) \le \eta_n\right\} - \gamma_n \mathsf{P}_1^{\otimes n}\left\{\sum_{i=1}^{n} \mathrm{LLR}(X_i) = \eta_n\right\}$$

Asymptotic performance

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

To highlight the role of each LLR, let $L_i \equiv \mathrm{LLR}(X_i)$. $L_1, ..., L_n$ are also i.i.d., according to distribution $\mathsf{Q}_\theta \equiv \mathsf{P}_\theta \circ \mathrm{LLR}^{-1}$.

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

$$\pi_{1|0}^{(n)} \approx \mathsf{Q}_0^{\otimes n}\left\{\sum_{i=1}^n L_i \geq \eta_n\right\}, \quad \pi_{0|1}^{(n)} \approx \mathsf{Q}_1^{\otimes n}\left\{\sum_{i=1}^n L_i \leq \eta_n\right\}.$$

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

- $\eta_n = n(\mathsf{E}_{X\sim\mathsf{P}_0}[\log \frac{\mathsf{P}_1(X)}{\mathsf{P}_0(X)}] + \delta)$ for some $\delta > 0$, and
- $\eta_n = n(\mathsf{E}_{X\sim\mathsf{P}_1}[\log \frac{\mathsf{P}_1(X)}{\mathsf{P}_0(X)}] - \delta')$ for some $\delta' > 0$,

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

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 $\eta_n \approx n\mathsf{E}_{X\sim \mathsf{P}_0}[\log \frac{\mathsf{P}_1(X)}{\mathsf{P}_0(X)}] =: -n\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)$.

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

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

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

$$\lim_{n\to\infty} -\frac{1}{n}\log \varpi_{0|1}^*(n,\epsilon) = \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)\,.$$

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

$$\lim_{n\to\infty} -\frac{1}{n}\log \varpi^*_{0|1}(n,\epsilon) = \mathsf{E}_{X\sim \mathsf{P}_0}\left[\log \frac{\mathsf{P}_0(X)}{\mathsf{P}_1(X)}\right] =: \mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)\,,$$

that is, $\varpi^*_{0|1}(n,\epsilon) = 2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+o(1))}$, or equivalently written as

$$\varpi^*_{0|1}(n,\epsilon) \doteq 2^{-n\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)}.$$

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 $\mathsf{P}_1^{\otimes n}$ to $\mathsf{P}_0^{\otimes n}$ and develop an exponentially tight lower bound.

Binary hypothesis testing

$$\begin{aligned} \underset{\mathcal{A}\subseteq\mathcal{X}^n}{\text{minimize}} \quad & \mathsf{P}_1^{\otimes n}(\mathcal{A}) \\ \text{subject to} \quad & \mathsf{P}_0^{\otimes n}(\mathcal{A}) \geq 1-\epsilon \end{aligned}$$

Lossless source coding

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

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

$$\mathcal{A} = \left\{ x^n \,\middle|\, \tfrac{1}{n}\sum_{i=1}^n f(x_i) \approx \mathsf{E}_{\mathsf{P}}[f(X)] \right\}.$$

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

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

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

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

Binary hypothesis testing

$$\begin{aligned} \underset{\mathcal{A}\subseteq\mathcal{X}^n}{\text{minimize}} \quad & \mathsf{P}_1^{\otimes n}(\mathcal{A}) \\ \text{subject to} \quad & \mathsf{P}_0^{\otimes n}(\mathcal{A}) \geq 1-\epsilon \end{aligned}$$

Lossless source coding

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

Now in the binary hypothesis testing problem, we again want to leverage the WLLN to ensure that the constraint $\mathsf{P}_0^{\otimes n}(\mathcal{A}) \geq 1-\epsilon$, where

$$\mathcal{A} = \left\{ x^n \,\middle|\, \tfrac{1}{n}\sum_{i=1}^n f(x_i) \approx \mathsf{E}_{\mathsf{P}_0}[f(X)] \right\}.$$

To control $\mathsf{P}_1^{\otimes n}(\mathcal{A})$ with $\mathsf{P}_0^{\otimes n}(\mathcal{A}) \approx 1$, now we should instead control $\mathsf{P}_0^{\otimes n}(x^n)/\mathsf{P}_1^{\otimes n}(x^n)$ for any sequence $x^n \in \mathcal{A}$.

Hence, we pick $f : x \mapsto \log\left(\mathsf{P}_0(x)/\mathsf{P}_1(x)\right)$ so that $\forall\, x^n \in \mathcal{A}$,

$$\mathsf{P}_0^{\otimes n}(x^n)/\mathsf{P}_1^{\otimes n}(x^n) \approx 2^{n\mathsf{E}_{\mathsf{P}_0}[\log \frac{\mathsf{P}_0(X)}{\mathsf{P}_1(X)}]}.$$

Consequently, $\mathsf{P}_1^{\otimes n}(\mathcal{A}) \approx 1/2^{n\mathsf{E}_{\mathsf{P}_0}[\log \frac{\mathsf{P}_0(X)}{\mathsf{P}_1(X)}]} = 2^{-n\mathsf{E}_{\mathsf{P}_0}[\log \frac{\mathsf{P}_0(X)}{\mathsf{P}_1(X)}]}$.

Typicality

Definition 5

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

$$\mathcal{A}^{(n)}(\mathsf{P}\|\mathsf{Q}) := \left\{ x^n \in \mathcal{X}^n \;\middle|\; \left| \tfrac{1}{n} \sum_{i=1}^n \log \tfrac{\mathsf{P}(x_i)}{\mathsf{Q}(x_i)} - \mathrm{D}(\mathsf{P}\|\mathsf{Q}) \right| \le \delta \right\}$$

Proposition 3

1. $\forall\, x^n \in \mathcal{A}_\delta^{(n)}(\mathsf{P}\|\mathsf{Q})$, $2^{n(\mathrm{D}(\mathsf{P}\|\mathsf{Q})-\delta)} \le \mathsf{P}^{\otimes n}(x^n)/\mathsf{Q}^{\otimes n}(x^n) \le 2^{n(\mathrm{D}(\mathsf{P}\|\mathsf{Q})+\delta)}$.
2. $\mathsf{P}^{\otimes n}\left(\mathcal{A}_\delta^{(n)}(\mathsf{P}\|\mathsf{Q})\right) \ge 1-\epsilon$ *for n large enough.*
3. $\mathsf{Q}^{\otimes n}\left(\mathcal{A}_\delta^{(n)}(\mathsf{P}\|\mathsf{Q})\right) \le 2^{-n(\mathrm{D}(\mathsf{P}\|\mathsf{Q})-\delta)}$.
4. $\mathsf{Q}^{\otimes n}\left(\mathcal{A}_\delta^{(n)}(\mathsf{P}\|\mathsf{Q})\right) \ge (1-\epsilon)2^{-n(\mathrm{D}(\mathsf{P}\|\mathsf{Q})+\delta)}$ *for n large enough.*

Achievability of Theorem 2

pf: Our goal is to show that there exist a sequence of tests $\left\{\phi^{(n)}\right\}$ with $\pi_{1|0}^{(n)} \leq \epsilon$ and $\forall\, \delta > 0$, $\pi_{0|1}^{(n)} \leq 2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\delta)}$ when n is large enough.

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

$$\phi^{(n)}(x^n) = \begin{cases} 0, & \text{if } x^n \in \mathcal{A}_\delta^{(n)}(\mathsf{P}_0\|\mathsf{P}_1) \\ 1, & \text{otherwise} \end{cases}$$

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,

$$\pi_{0|1}^{(n)}(\phi^{(n)}) = \mathsf{P}_1^{\otimes n}\left(\mathcal{A}_\delta^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)\right) \leq 2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)-\delta)}.$$

This completes the proof of the achievability. □

Converse of Theorem 2

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

$$\begin{aligned}\pi_{0|1}^{(n)}(\phi^{(n)}) &= \mathsf{E}_{\mathsf{P}_1^{\otimes n}}\left[1-\phi^{(n)}(X^n)\right] \\ &\ge \mathsf{E}_{\mathsf{P}_1^{\otimes n}}\left[\left(1-\phi^{(n)}(X^n)\right)\mathbb{1}\left\{X^n \in \mathcal{A}_{\delta'}^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)\right\}\right], \end{aligned} \tag{3}$$

where the last step is taking the expectation over the typical set $\mathcal{A}_{\delta'}^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)$.

Next, to leverage the fact that $\pi_{1|0}^{(n)}(\phi^{(n)}) = \mathsf{E}_{\mathsf{P}_0^{\otimes n}}\left[\phi^{(n)}(X^n)\right] \le \epsilon$ to further lower bound (3), we would like to change the distribution in the expectation from $\mathsf{P}_1^{\otimes n}$ to $\mathsf{P}_0^{\otimes n}$. That is, for each $x^n \in \mathcal{A}_{\delta'}^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)$ in the expectation, we change $\mathsf{P}_1^{\otimes n}(x^n)$ to $\mathsf{P}_0^{\otimes n}(x^n)$. By property 1 of Proposition 3, change of measure leads to a further lower bound:

$$(3) \ge \mathsf{E}_{\mathsf{P}_0^{\otimes n}}\left[\left(1-\phi^{(n)}(X^n)\right)\mathbb{1}\left\{X^n \in \mathcal{A}_{\delta'}^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)\right\}\right] 2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\delta')}. \tag{4}$$

Further observe that

$$\begin{aligned}
&\mathsf{E}_{\mathsf{P}_0^{\otimes n}}\left[\left(1-\phi^{(n)}(X^n)\right)\mathbb{1}\left\{X^n\in\mathcal{A}_{\delta'}^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)\right\}\right]\\
&=\mathsf{P}_0^{\otimes n}\left(\mathcal{A}_{\delta'}^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)\right)-\mathsf{E}_{\mathsf{P}_0^{\otimes n}}\left[\phi^{(n)}(X^n)\mathbb{1}\left\{X^n\in\mathcal{A}_{\delta'}^{(n)}(\mathsf{P}_0\|\mathsf{P}_1)\right\}\right]\\
&\geq(1-\epsilon')-\mathsf{E}_{\mathsf{P}_0^{\otimes n}}\left[\phi^{(n)}(X^n)\right]\quad\text{for } n \text{ large enough.}
\end{aligned}\tag{5}$$

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

Combining (3) – (5), and the fact that $\pi_{1|0}^{(n)}(\phi^{(n)})=\mathsf{E}_{\mathsf{P}_0^{\otimes n}}\left[\phi^{(n)}(X^n)\right]\leq\epsilon$, we arrive at

$$\pi_{0|1}^{(n)}(\phi^{(n)})\geq(1-\epsilon'-\epsilon)\,2^{-n(\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)+\delta')}\quad\text{for } n \text{ large enough.}$$

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.

- $\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)$ 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.

  $\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)$ 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 $\mathrm{LR}(x) = \frac{P_1(x)}{P_0(x)}$ 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:

$$\varpi^*_{0|1}(n, \epsilon) \doteq 2^{-n\mathrm{D}(P_0 \| P_1)}.$$

Next: definition, interpretation, and properties of information divergence

Information divergence

Recap: Consider the task of determining from i.i.d. sampled data $X_1, ..., X_n$ 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

$$\varpi^*_{0|1}(n, \epsilon) \doteq 2^{-n\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)}.$$

So, getting one more sample, roughly speaking, the probability of error in the logarithmic scale decrease by an amount $\mathrm{D}(\mathsf{P}_0\|\mathsf{P}_1)$. 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

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

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

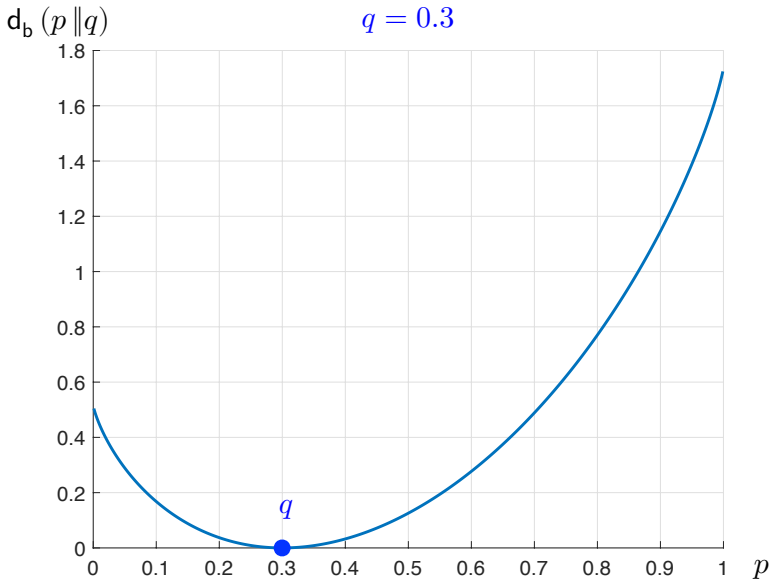
  Note: if $\mathsf{supp}\mathsf{P}\setminus\mathsf{supp}\mathsf{Q}\neq\emptyset$, then $\mathrm{D}(\mathsf{P}\|\mathsf{Q}) = +\infty$.

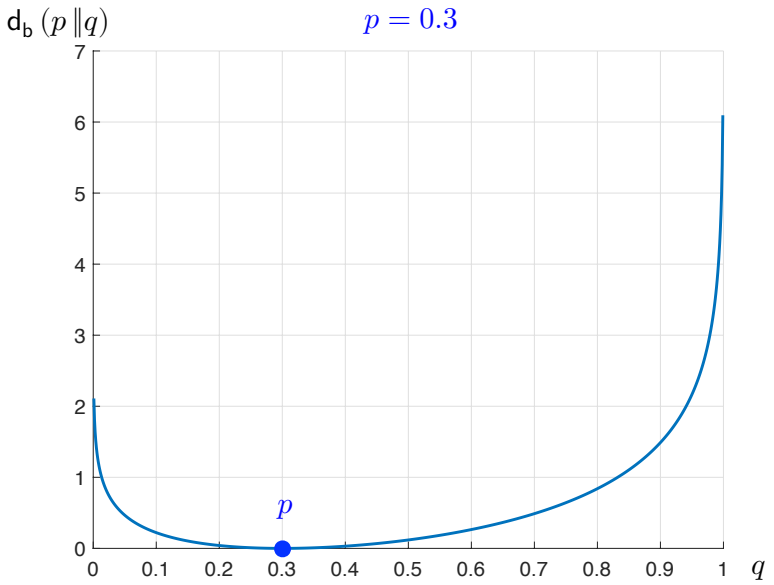
- $\mathcal{A} = \mathbb{R}^d$ and P and Q are PDF's: if $\mathsf{supp}\mathsf{P}\setminus\mathsf{supp}\mathsf{Q}$ has zero volume (Lebesgue measure = 0), then

$$\mathrm{D}(\mathsf{P}\|Q) := \int_{\mathbb{R}^d} \mathsf{P}(\boldsymbol{x})\log\frac{\mathsf{P}(\boldsymbol{x})}{\mathsf{Q}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} = \int_{\mathsf{supp}\mathsf{P}} \mathsf{P}(\boldsymbol{x})\log\frac{\mathsf{P}(\boldsymbol{x})}{\mathsf{Q}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x}.$$

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





Example 1 (Binary divergence function)

Let $\boldsymbol{p} = (1-p, p)$ and $\boldsymbol{q} = (1-q, q)$, $p, q \in (0,1)$. The information divergence of $\boldsymbol{p}$ from $\boldsymbol{q}$ is called the binary divergence function $\mathsf{d}_\mathsf{b}(p\|q)$, where

$$\begin{aligned}\mathsf{d}_\mathsf{b}(p\|q) &:= \mathrm{D}(\boldsymbol{p}\|\boldsymbol{q}) \\ &= p\log\frac{p}{q} + (1-p)\log\frac{(1-p)}{(1-q)}.\end{aligned}$$

Exercise 1

Prove the following statements.

1. $\mathsf{d}_\mathsf{b}\left(p\left\|\frac{1}{2}\right.\right) = 1 - \mathsf{H}_\mathsf{b}(p)$.
2. $\mathsf{d}_\mathsf{b}(p\|q)$ is convex in (p, q).
3. $\mathsf{d}_\mathsf{b}(p\|q) \geq 2(p-q)^2\log(e)$.

Non-negativity of information divergence

Theorem 3

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

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

$$-\mathrm{D}(\mathsf{P}\|\mathsf{Q}) = \mathsf{E}_{X\sim\mathsf{P}}\left[\log \frac{\mathsf{Q}(X)}{\mathsf{P}(X)}\right]$$

$$\overset{\text{Jensen}}{\leq} \log \mathsf{E}_{X\sim\mathsf{P}}\left[\frac{\mathsf{Q}(X)}{\mathsf{P}(X)}\right] = \log \sum_{a\in\mathsf{supp}\mathsf{P}} \mathsf{Q}(a) \leq \log 1 = 0,$$

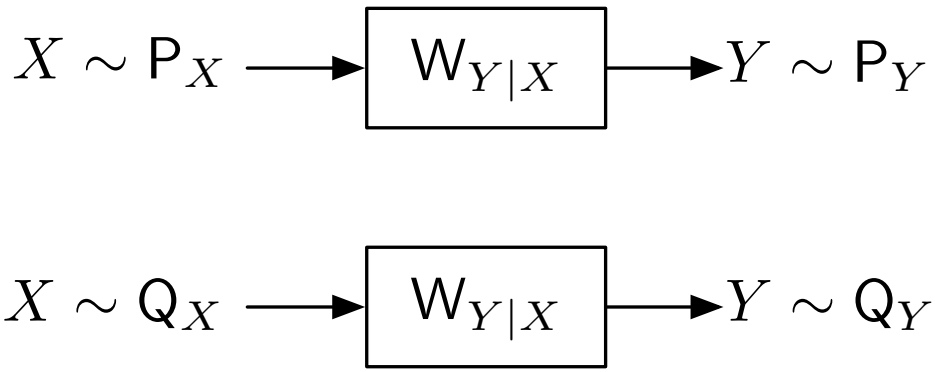
where the first inequality is due to the fact that $x \mapsto \log x$ is strictly concave.

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

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

Data processing

We always process data to extract information.



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* $\mathsf{W}_{Y|X}(y|x)$ 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)

$$\mathrm{D}(\mathsf{P}_X \| \mathsf{Q}_X) \geq \mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y).$$

pf: Let the joint laws of (X, Y) in the upper branch and the lower branch be $\mathsf{P}_{X,Y}$ and $\mathsf{Q}_{X,Y}$ respectively.

Nota that $\forall\, x, y$, $\frac{\mathsf{P}_{X,Y}(x,y)}{\mathsf{Q}_{X,Y}(x,y)} = \frac{\mathsf{P}_X(x)\mathsf{W}_{Y|X}(y|x)}{\mathsf{Q}_X(x)\mathsf{W}_{Y|X}(y|x)} = \frac{\mathsf{P}_X(x)}{\mathsf{Q}_X(x)}$. Hence,

$$\begin{aligned}
-\mathrm{D}(\mathsf{P}_X \| \mathsf{Q}_X) &= -\mathrm{D}(\mathsf{P}_{X,Y} \| \mathsf{Q}_{X,Y}) \\
&= \mathsf{E}_{Y\sim\mathsf{P}_Y}\left[\mathsf{E}_{X\sim\mathsf{P}_{X|Y}}\left[\log \frac{\mathsf{Q}_{X,Y}(X,Y)}{\mathsf{P}_{X,Y}(X,Y)}\middle| Y\right]\right] \\
&\leq \mathsf{E}_{Y\sim\mathsf{P}_Y}\left[\log \mathsf{E}_{X\sim\mathsf{P}_{X|Y}}\left[\frac{\mathsf{Q}_{X,Y}(X,Y)}{\mathsf{P}_{X,Y}(X,Y)}\middle| Y\right]\right] \\
&= \mathsf{E}_{Y\sim\mathsf{P}_Y}\left[\log \frac{\mathsf{Q}_Y(Y)}{\mathsf{P}_Y(Y)}\right] = -\mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y)\,,
\end{aligned}$$

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 $\mathsf{P}_{Y|X}$ and $\mathsf{Q}_{Y|X}$ be two conditional laws of Y given X and P_X be a marginal law of X. The conditional information divergence of $\mathsf{P}_{Y|X}$ from $\mathsf{Q}_{Y|X}$ conditioned on P_X is defined as

$$\mathrm{D}\left(\mathsf{P}_{Y|X}\middle\|\mathsf{Q}_{Y|X}\middle|\mathsf{P}_X\right) := \mathsf{E}_{X\sim\mathsf{P}_X}\left[\mathrm{D}\left(\mathsf{P}_{Y|X}(\cdot|X)\middle\|\mathsf{Q}_{Y|X}(\cdot|X)\right)\right].$$

Exercise 2 (Nonnegativity of Conditional Divergence)

Show $\mathrm{D}\left(\mathsf{P}_{Y|X}\middle\|\mathsf{Q}_{Y|X}\middle|\mathsf{P}_X\right) \geq 0$.

Chain rule for information divergence

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

Theorem 5 (Chain Rule for Information Divergence)

Let $\mathsf{P}_{X,Y}$ *and* $\mathsf{Q}_{X,Y}$ *be two joint laws of* (X,Y) *with conditional laws of* Y *given* X *being* $\mathsf{P}_{Y|X}$ *and* $\mathsf{Q}_{Y|X}$ *respectively and marginal laws of* X *being* P_X *and* Q_X *respectively. Then,*

$$\mathrm{D}(\mathsf{P}_{X,Y}\|\mathsf{Q}_{X,Y}) = \mathrm{D}\left(\mathsf{P}_{Y|X}\middle\|\mathsf{Q}_{Y|X}\middle|\mathsf{P}_X\right) + \mathrm{D}(\mathsf{P}_X\|\mathsf{Q}_X)\,.$$

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

Exercise 3

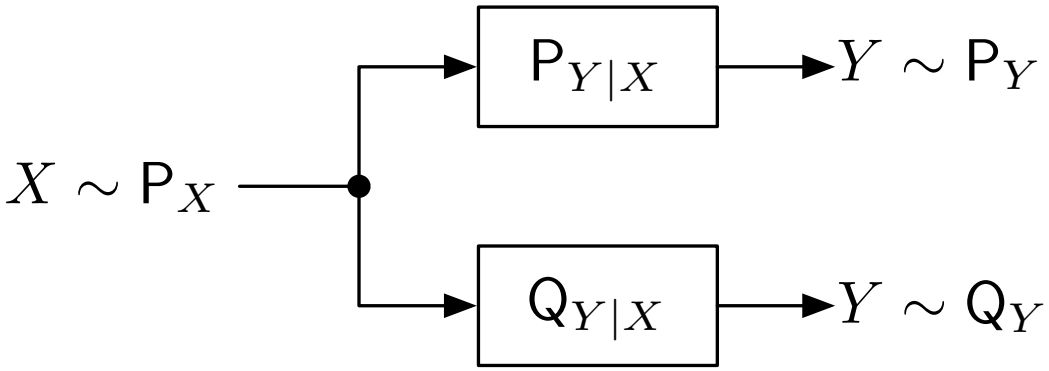
Let P_{X^n} and Q_{X^n} be joint laws of X^n with conditional laws and marginal laws defined similarly as above. Show that

$$\mathrm{D}(\mathsf{P}_{X^n}\|\mathsf{Q}_{X^n}) = \sum_{i=1}^{n} \mathrm{D}\left(\mathsf{P}_{X_i|X^{i-1}}\middle\|\mathsf{Q}_{X_i|X^{i-1}}\middle|\mathsf{P}_{X^{i-1}}\right).$$

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

$$\mathrm{D}(\mathsf{P}_{X^n}\|\mathsf{Q}_{X^n}) = \mathrm{D}\left(\mathsf{P}_{X^n}\middle\|\prod_{i=1}^{n}\mathsf{P}_{X_i}\right) + \sum_{i=1}^{n}\mathrm{D}(\mathsf{P}_{X_i}\|\mathsf{Q}_{X_i}).$$

Conditioning increases information divergence



Theorem 6 (Conditioning Increases Information Divergence)

Let $\mathsf{P}_{Y|X}$ *and* $\mathsf{Q}_{Y|X}$ *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* $X \sim \mathsf{P}_X$ *through* $\mathsf{P}_{Y|X}$ *and* $\mathsf{Q}_{Y|X}$ *respectively. Then,*

$$\mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y) \le \mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{Q}_{Y|X} \middle| \mathsf{P}_X\right), \text{ with equality iff } \mathrm{D}\left(\mathsf{P}_{X|Y} \middle\| \mathsf{Q}_{X|Y} \middle| \mathsf{P}_Y\right) = 0.$$

Conditioned on the common "input" law, dissimilarity between $\mathsf{P}_{Y|X}$ and $\mathsf{Q}_{Y|X}$ 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 $\mathsf{P}_{X,Y}$ and $\mathsf{Q}_{X,Y}$ respectively.

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

By the Chain Rule (Theorem 5),

$$\mathrm{D}(\mathsf{P}_{X,Y}\|\mathsf{Q}_{X,Y}) = \mathrm{D}\left(\mathsf{P}_{Y|X}\middle\|\mathsf{Q}_{Y|X}\middle|\mathsf{P}_X\right) + \mathrm{D}(\mathsf{P}_X\|\mathsf{P}_X) = \mathrm{D}\left(\mathsf{P}_{Y|X}\middle\|\mathsf{Q}_{Y|X}\middle|\mathsf{P}_X\right)$$
$$\mathrm{D}(\mathsf{P}_{X,Y}\|\mathsf{Q}_{X,Y}) = \mathrm{D}\left(\mathsf{P}_{X|Y}\middle\|\mathsf{Q}_{X|Y}\middle|\mathsf{P}_Y\right) + \mathrm{D}(\mathsf{P}_Y\|\mathsf{Q}_Y)$$

The proof is complete by comparing the above two and the non-negativity of $\mathrm{D}\left(\mathsf{P}_{X|Y}\middle\|\mathsf{Q}_{X|Y}\middle|\mathsf{P}_Y\right)$. □

Convexity of information divergence

Theorem 7

$\mathrm{D}(\mathsf{P}\|\mathsf{Q})$ *is convex in* (P,Q)*, that is, for any* $\lambda\in[0,1]$ *and* $\mathsf{P}_0,\mathsf{P}_1,\mathsf{Q}_0,\mathsf{Q}_1$*,*

$$\mathrm{D}(\mathsf{P}_\lambda\|\mathsf{Q}_\lambda)\le(1-\lambda)\mathrm{D}(\mathsf{P}_0\|\mathsf{Q}_0)+\lambda\mathrm{D}(\mathsf{P}_1\|\mathsf{Q}_1)\,,$$

where $\mathsf{P}_\lambda=(1-\lambda)\mathsf{P}_0+\lambda\mathsf{P}_1$ *and* $\mathsf{Q}_\lambda=(1-\lambda)\mathsf{Q}_0+\lambda\mathsf{Q}_1$*.*

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* $\mathrm{D}(\mathsf{P}\|\mathsf{Q}) < \infty$*, then for any* $f : \mathcal{X} \to \mathbb{R}$ *such that* $\mathsf{E}_{X\sim\mathsf{Q}}[2^{f(X)}] < \infty$*,* $\mathsf{E}_{X\sim\mathsf{P}}[f(X)]$ *exists. Furthermore,*

$$\mathrm{D}(\mathsf{P}\|\mathsf{Q}) = \max_{f:\mathcal{X}\to\mathbb{R}} \mathsf{E}_{X\sim\mathsf{P}}[f(X)] - \log \mathsf{E}_{X\sim\mathsf{Q}}[2^{f(X)}] \quad \textit{s.t. } \mathsf{E}_{X\sim\mathsf{Q}}[2^{f(X)}] < \infty.$$

pf: There are two directions to prove: "$\leq$" and "$\geq$".

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

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

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

$$\mathsf{E}_{X\sim\mathsf{P}}\Bigg[\log \mathsf{P}(X) - \underbrace{\Big(f(X) + \log \mathsf{Q}(X) - \log \mathsf{E}_{X\sim\mathsf{Q}}[2^{f(X)}]\Big)}_{(*)}\Bigg] \geq 0.$$

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

The proof is complete by $\mathrm{D}(\mathsf{P}\|\mathsf{Q}_f) \geq 0$. □

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.