Hypothesis Testing and KL divergence— Lectures 6, 7

Haruka Eshima

October 2026

Abstract

This is a lecture note that I took for the Information Theory class taught by Prof. Amedeo Roberto Esposito at the Okinawa Institute of Science and Technology, 2026. Any errors are my own.
The questions in green are made by me and are not officially related to the mini test in the class.

Convention on logarithms. $\log = \log_2$ throughout, so entropies and divergences are in bits and exponentials are written $2^{(\cdot)}$. The handwritten notes switch to $e^{(\cdot)}$ in the divergence-typical-set section; see the correction there. Conventions for the divergence: $0 \log \frac{0}{p} = 0$ for $p \in [0, 1]$ and $q \log \frac{q}{0} = +\infty$ for $q > 0$.

Contents

Summary **2**

1 Where the divergence comes from: the pmf induced by a code **2**

2 Universal compression **3**

3 Compression is prediction **4**

4 The general definition and six properties **5**
4.1 Property 1: non-negativity . . . 5
4.2 Property 2: asymmetry, and failure of the triangle inequality . . . 5
4.3 Property 3: unboundedness . . . 6
4.4 Property 4: relation to entropy . . . 6
4.5 Property 5: chain rule and tensorisation . . . 6
4.6 Property 6: data processing inequality . . . 6

5 Hypothesis testing **7**
5.1 The Neyman–Pearson lemma . . . 8

6 Divergence typical set **8**

7 Stein's lemma **9**
7.1 Achievability (upper bound on β_n^*) . . . 9
7.2 Optimality (lower bound on β_n^*) . . . 10

Summary

- For every pmf $Q \gg P$ (i.e. $Q(x) = 0 \Rightarrow P(x) = 0$),
$$D\big(P \,\|\, Q\big) + H(P) \;\le\; \mathbb{E}_P\big[\ell_{\mathscr{C}_Q}(X)\big] \;<\; D\big(P \,\|\, Q\big) + H(P) + 1,$$
where $\mathscr{C}_Q$ is the Shannon code built for Q.
- For every prefix-free code $\mathscr{C}$ on $\mathcal{X}$, with $Q_{\mathscr{C}}$ the pmf induced by $\mathscr{C}$ via $Q_{\mathscr{C}}(x) = 2^{-\ell_{\mathscr{C}}(x)}$,
$$\mathbb{E}_P\big[\ell_{\mathscr{C}}(X)\big] \;=\; D\big(P \,\|\, Q_{\mathscr{C}}\big) + H(P).$$
- Six properties of KL divergence, e.g. the data processing inequality.
- Hypothesis testing and the two types of error.
- Neyman–Pearson lemma.
- Divergence typical set.
- Stein's lemma.

1 Where the divergence comes from: the pmf induced by a code

Let $X \sim P$ and let $\mathscr{C}$ be a prefix-free code for the DMS X.

Definition 1.1 (pmf induced by a code)**.** Define
$$Q(x) := 2^{-\ell_{\mathscr{C}}(x)}, \qquad x \in \mathcal{X}.$$
By Kraft's inequality $\sum_{x\in\mathcal{X}} Q(x) \le 1$, so Q need not be a pmf on $\mathcal{X}$. Adjoin a dummy symbol $*$ and set
$$Q(*) := 1 - \sum_{x\in\mathcal{X}} Q(x) \;\ge\; 0.$$
Then Q is a genuine pmf on $\mathcal{X} \cup \{*\}$, and P — extended by $P(*) := 0$ — satisfies $Q \gg P$.

How efficient is $\mathscr{C}$?

Proposition 1.2. *With Q as in Definition 1.1,*
$$\mathbb{E}_P\big[\ell_{\mathscr{C}}(X)\big] - H(X) = \sum_x P(x)\Big(\log \tfrac{1}{Q(x)} - \log \tfrac{1}{P(x)}\Big) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = D\big(P \,\|\, Q\big) \;\ge\; 0.$$

The excess of a code's expected length over the entropy is exactly the divergence between the true source and the distribution the code "believes in". The term $*$ contributes nothing because $P(*) = 0$ and $0 \log \frac{0}{Q(*)} = 0$.

Definition 1.3 (Kullback–Leibler divergence)**.** Let P, Q be two discrete probability distributions on the same set $\mathcal{X}$. The *Kullback–Leibler divergence* between P and Q is
$$D\big(P \,\|\, Q\big) := \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)},$$
with the conventions $0 \log \frac{0}{p} = 0$ for every $p \in [0, 1]$ and $q \log \frac{q}{0} = +\infty$ for every $q > 0$.

The code induced by a distribution

Conversely, given a pmf Q on $\mathcal{X}$, define the *Shannon code for* Q by

$$\ell_{\mathscr{C}_Q}(x) := \left\lceil \log \tfrac{1}{Q(x)} \right\rceil.$$

This satisfies Kraft, so such a prefix-free code exists. Then, for the true source $X \sim P$,

$$D\big(P \,\|\, Q\big) = \mathbb{E}_P\Big[\log \tfrac{P(X)}{Q(X)}\Big] \;\le\; \mathbb{E}_P\big[\ell_{\mathscr{C}_Q}(X)\big] - H(X) \;\le\; \mathbb{E}_P\Big[\log \tfrac{1}{Q(X)} + 1 + \log P(X)\Big] = D\big(P \,\|\, Q\big) + 1,$$

hence

$$H(X) + D\big(P \,\|\, Q\big) \le \mathbb{E}_P\big[\ell_{\mathscr{C}_Q}(X)\big] \;\le\; H(X) + D\big(P \,\|\, Q\big) + 1.$$

Remark 1.4. If $P = Q$ this recovers $\mathbb{E}[\ell] \le H(X) + 1$, i.e. the Shannon code built for the *correct* distribution is efficient. **The divergence $D\big(P \,\|\, Q\big)$ is exactly the price in bits per symbol of designing the code for the wrong distribution.**

2 Universal compression

Now suppose we only partially know the source: $P \in S \subseteq \Pi(\mathcal{X})$, the simplex of pmfs on $\mathcal{X}$. We want to pick

$$Q^* \in \arg\min_{Q} \sup_{P \in S} D\big(P \,\|\, Q\big),$$

the minimax-redundancy distribution.

Definition 2.1 (Universal / asymptotically optimal code)**.** A sequence of prefix-free codes $(\mathscr{C}_n)_{n \ge 1}$, $\mathscr{C}_n : \mathcal{X}^n \to \{0,1\}^*$, is *universal* (or *asymptotically optimal*) for a class S of sources if its per-symbol redundancy vanishes for *every* source in the class:

$$\forall P \in S: \qquad \lim_{n \to \infty} \Big(\tfrac{1}{n} \mathbb{E}_{P^n}\big[\ell_{\mathscr{C}_n}(X^n)\big] - H(P) \Big) = 0.$$

The point is that a single sequence of codes — which does not know which $P \in S$ generated the data — achieves the entropy rate of each of them.

Example 2.2 (Binary i.i.d. sources; the Laplace estimator)**.** Let $X^n \overset{\text{iid}}{\sim} \text{Bern}(\theta)$ and

$$S_n = \big\{P_\theta^n : 0 < \theta < 1\big\}, \qquad X = \begin{cases} 1 & \text{w.p. } \theta \\ 0 & \text{w.p. } 1 - \theta. \end{cases}$$

The naive estimate fails. The naive sequential estimate is

$$\widehat{Q}(1 \mid x^k) = \frac{\#1\text{s in } x^k}{k}.$$

But if $x_1 = 0$ then $\widehat{Q}(1 \mid x_1) = 0$, so $\widehat{Q}^{\,2}(01) = 0$ while $P_\theta^2(01) = (1-\theta)\theta > 0$; hence $D\big(P_\theta^2 \,\|\, \widehat{Q}^{\,2}\big) = +\infty$ whenever $\theta > 0$.

Modification (add-one / Laplace). Mix the empirical frequency with the uniform distribution:

$$Q(1 \mid x^k) = \frac{\#1\text{s in } x^k}{k} \cdot \frac{k}{k+2} + \frac{1}{2} \cdot \frac{2}{k+2} = \frac{n_1(x^k) + 1}{k+2}, \qquad Q(0) = Q(1) = \tfrac{1}{2}.$$

By induction, writing $n_0(x^n), n_1(x^n)$ for the counts in x^n,

$$Q(x^n) \;\ge\; \frac{1}{n+1} \left(\frac{n_0(x^n)}{n}\right)^{n_0(x^n)} \left(\frac{n_1(x^n)}{n}\right)^{n_1(x^n)}.$$

The redundancy is at most $\log(n+1)$

Using the bound of Example 2.2: since $\log P_\theta^n(x^n) = n_0 \log(1-\theta) + n_1 \log\theta$,

$$\begin{aligned} D\big(P_\theta^n \,\big\|\, Q\big) &= \mathbb{E}_{P_\theta^n}\Big[\log \frac{P_\theta^n(X^n)}{Q(X^n)}\Big] \\ &\le \log(n+1) + n\,\mathbb{E}\Big[\tfrac{n_0(X^n)}{n} \log \tfrac{1-\theta}{n_0(X^n)/n} + \tfrac{n_1(X^n)}{n} \log \tfrac{\theta}{n_1(X^n)/n}\Big] \\ &= \log(n+1) - n\,\mathbb{E}\Big[D\Big(\big(\tfrac{n_0}{n}, \tfrac{n_1}{n}\big) \,\Big\|\, (1-\theta,\theta)\Big)\Big] \;\le\; \log(n+1), \end{aligned}$$

the last step because the divergence inside the expectation is non-negative.

Consequently, for the induced code $\mathscr{C}_Q^n$,

$$\frac{1}{n}\mathbb{E}\big[\ell_{\mathscr{C}_Q^n}(X^n)\big] - \frac{1}{n}H(X^n) \;<\; \frac{1}{n}\Big(D\big(P_\theta^n \,\big\|\, Q\big) + 1\Big) \;<\; \frac{1}{n}\big(\log(n+1)+1\big) \;\longrightarrow\; 0 \qquad (n\to\infty),$$

so $(\mathscr{C}_Q^n)_n$ is universal for S in the sense of Definition 2.1.

3 Compression is prediction

Suppose $P \in S$ and we have a code $\mathscr{C} : \mathcal{X}^n \to \{0,1\}^*$ that is "good" uniformly over S: for some $\varepsilon > 0$,

$$\forall P \in S: \qquad \frac{1}{n}\mathbb{E}\big[\ell_{\mathscr{C}}(X^n)\big] - H(X) < \varepsilon.$$

By Proposition 1.2 applied at block length n, with Q the pmf on $\mathcal{X}^n$ induced by $\mathscr{C}$,

$$\frac{1}{n}D\big(P^n \,\big\|\, Q\big) \;\le\; \frac{1}{n}\mathbb{E}\big[\ell_{\mathscr{C}}(X^n)\big] - H(X) \;<\; \varepsilon.$$

Expanding by the chain rule:

$$\begin{aligned} \varepsilon > \frac{1}{n}D\big(P^n \,\big\|\, Q\big) &= \frac{1}{n}\sum_{x^n} P(x^n) \log \frac{P(x^n)}{Q(x^n)} \\ &= \frac{1}{n}\sum_{x^n} P(x^n) \log \frac{P(x_n \mid x^{n-1})P(x_{n-1} \mid x^{n-2}) \cdots P(x_2 \mid x_1)P(x_1)}{Q(x_n \mid x^{n-1})Q(x_{n-1} \mid x^{n-2}) \cdots Q(x_2 \mid x_1)Q(x_1)} \\ &= \frac{1}{n}\sum_{i=1}^{n} \sum_{x^i} P(x^{i-1})P(x_i \mid x^{i-1}) \log \frac{P(x_i \mid x^{i-1})}{Q(x_i \mid x^{i-1})} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{X^{i-1}}}\Big[D\big(P(\cdot \mid X^{i-1}) \,\big\|\, Q(\cdot \mid X^{i-1})\big)\Big]. \end{aligned}$$

Remark 3.1. Knowing how to compress x^n is equivalent to knowing how to predict the next symbol given x^{i-1}: a code with small redundancy is exactly a predictor whose conditional distributions are close, on average, to the true ones. Cf. the Shannon game and the claim that "the entropy of English text is small".

4 The general definition and six properties

Let P, Q be probability measures and assume $P \ll Q$ (absolute continuity: $Q(A) = 0 \Rightarrow P(A) = 0$ for every A in the σ-algebra). Then there is a non-negative measurable $f := \frac{dP}{dQ}$ (the Radon–Nikodym derivative) with

$$P(A) = \int_A f(x)\, Q(dx) \qquad \text{for every measurable } A,$$

and

$$D\big(P \,\|\, Q\big) = \mathbb{E}_Q\Big[\tfrac{dP}{dQ} \log \tfrac{dP}{dQ}\Big].$$

The idea is the same as a change of measure.

Q. *State and prove the six properties of KL divergence, including the data processing inequality.*

4.1 Property 1: non-negativity

Proposition 4.1. $D\big(P \,\|\, Q\big) \geq 0$, *with equality iff* $P = Q$.

Proof. Let $\psi(t) = t \log t$, which is strictly convex on $(0, \infty)$, and put $T := \frac{dP}{dQ}$. Then $\mathbb{E}_Q[T] = 1$ and by Jensen

$$D\big(P \,\|\, Q\big) = \mathbb{E}_Q\big[\psi(T)\big] \;\geq\; \psi\big(\mathbb{E}_Q[T]\big) = \psi(1) = 0.$$

By strict convexity, equality forces T to be Q-a.s. constant; since $\mathbb{E}_Q[T] = 1$ the constant is 1, i.e. $P = Q$. □

4.2 Property 2: asymmetry, and failure of the triangle inequality

Proposition 4.2. *In general* $D\big(P \,\|\, Q\big) \neq D\big(Q \,\|\, P\big)$, *and the triangle inequality can fail: there exist* P, W, Q *with*

$$D\big(P \,\|\, Q\big) \;>\; D\big(P \,\|\, W\big) + D\big(W \,\|\, Q\big).$$

Asymmetry. For small $\varepsilon > 0$,

$$D\big((\tfrac{1}{2}, \tfrac{1}{2}) \,\|\, (1-\varepsilon, \varepsilon)\big) = \tfrac{1}{2} \log \tfrac{1}{2(1-\varepsilon)} + \tfrac{1}{2} \log \tfrac{1}{2\varepsilon} \;\gg\; \log 2,$$

diverging as $\varepsilon \downarrow 0$, whereas

$$D\big((1-\varepsilon, \varepsilon) \,\|\, (\tfrac{1}{2}, \tfrac{1}{2})\big) = \log 2 + (1-\varepsilon)\log(1-\varepsilon) + \varepsilon \log \varepsilon \;\approx\; \log 2.$$

Counterexample to the triangle inequality. Take Bernoulli distributions $P = \mathrm{Bern}(0.4)$, $W = \mathrm{Bern}(0.5)$, $Q = \mathrm{Bern}(0.6)$. Then (in bits)

$$D\big(P \,\|\, W\big) \approx 0.0290, \qquad D\big(W \,\|\, Q\big) \approx 0.0294, \qquad D\big(P \,\|\, W\big) + D\big(W \,\|\, Q\big) \approx 0.0585,$$

while

$$D\big(P \,\|\, Q\big) = 0.4 \log \tfrac{0.4}{0.6} + 0.6 \log \tfrac{0.6}{0.4} \approx 0.117.$$

So $D\big(P \,\|\, Q\big) > D\big(P \,\|\, W\big) + D\big(W \,\|\, Q\big)$. □

▶ **Remark.** The mechanism is that KL behaves locally like a *squared* distance, and squared distances fail the triangle inequality for a midpoint ($1^2 > (\frac{1}{2})^2 + (\frac{1}{2})^2$).

4.3 Property 3: unboundedness

Proposition 4.3. *$D\big(P \,\big\|\, Q\big)$ can be made arbitrarily large, and is $+\infty$ when $P \not\ll Q$.*

Proof. Take $P = \mathrm{Bern}(1)$ (the point mass at 1) and $Q = \mathrm{Bern}(\varepsilon)$. Then

$$D\big(P \,\big\|\, Q\big) = 1 \cdot \log \frac{1}{\varepsilon} \;\longrightarrow\; +\infty \qquad \text{as } \varepsilon \downarrow 0.$$

□

4.4 Property 4: relation to entropy

Proposition 4.4. $D\big(P \,\big\|\, Q\big) = \mathbb{E}_P[-\log Q] - H(P)$.

Proof. $D\big(P \,\big\|\, Q\big) = \sum_x P(x) \log P(x) - \sum_x P(x) \log Q(x) = -H(P) + \mathbb{E}_P[-\log Q]$. □

4.5 Property 5: chain rule and tensorisation

Proposition 4.5.

$$D\big(P_{XY} \,\big\|\, Q_{XY}\big) = D\big(P_X \,\big\|\, Q_X\big) + \mathbb{E}_{X\sim P_X}\Big[D\big(P_{Y|X} \,\big\|\, Q_{Y|X}\big)\Big],$$

and for product measures

$$D\big(P_1 \otimes P_2 \,\big\|\, Q_1 \otimes Q_2\big) = D\big(P_1 \,\big\|\, Q_1\big) + D\big(P_2 \,\big\|\, Q_2\big), \qquad \textit{hence} \qquad D\big(P^{\otimes n} \,\big\|\, Q^{\otimes n}\big) = n\, D\big(P \,\big\|\, Q\big).$$

Proof.

$$\begin{aligned} D\big(P_{XY} \,\big\|\, Q_{XY}\big) &= \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{Q_{XY}(x,y)} \\ &= \sum_{x,y} P_{XY}(x,y) \Big(\log \frac{P_X(x)}{Q_X(x)} + \log \frac{P_{Y|X=x}(y)}{Q_{Y|X=x}(y)}\Big) \\ &= D\big(P_X \,\big\|\, Q_X\big) + \sum_x P_X(x)\, D\big(P_{Y|X=x} \,\big\|\, Q_{Y|X=x}\big). \end{aligned}$$

The product case is the special case $P_{Y|X} = P_2$, $Q_{Y|X} = Q_2$; the n-fold case follows by induction. □

4.6 Property 6: data processing inequality

Proposition 4.6 (DPI)**.** *Let $K = P_{Y|X}$ be a channel (a Markov kernel), and write $P_Y = PK$, $Q_Y = QK$ for the push-forwards of P_X and Q_X through K. Then*

$$D\big(PK \,\big\|\, QK\big) = D\big(P_Y \,\big\|\, Q_Y\big) \;\le\; D\big(P_X \,\big\|\, Q_X\big).$$

Proof. **Step 1: the likelihood ratio contracts to a conditional expectation.** Write

$$K_{Q,X|Y=y}(x) := \frac{Q_X(x)\, K_{Y|X=x}(y)}{Q_Y(y)}$$

for the *backward* channel induced by Q (a pmf in x for each y). Then

$$\frac{P_Y(y)}{Q_Y(y)} = \sum_x \frac{P_X(x)K_{Y|X=x}(y)}{Q_Y(y)} = \sum_x \frac{P_X(x)}{Q_X(x)} \cdot \underbrace{\frac{Q_X(x)K_{Y|X=x}(y)}{Q_Y(y)}}_{=K_{Q,X|Y=y}(x)} = \mathbb{E}_{K_{Q,X|Y=y}}\left[\frac{P_X}{Q_X}\right].$$

Step 2: Jensen. With $\psi(t) = t\log t$ convex,

$$\begin{aligned} D\big(P_Y \,\|\, Q_Y\big) &= \mathbb{E}_{Q_Y}\left[\tfrac{P_Y}{Q_Y}\log\tfrac{P_Y}{Q_Y}\right] = \mathbb{E}_{Q_Y}\left[\psi\Big(\mathbb{E}_{K_{Q,X|Y}}\left[\tfrac{P_X}{Q_X}\right]\Big)\right] \\ &\le \mathbb{E}_{Q_Y}\left[\mathbb{E}_{K_{Q,X|Y}}\left[\tfrac{P_X}{Q_X}\log\tfrac{P_X}{Q_X}\right]\right] = \mathbb{E}_{Q_{XY}}\left[\tfrac{P_X}{Q_X}\log\tfrac{P_X}{Q_X}\right]. \end{aligned}$$

Step 3: marginalise. Since the integrand depends on x only,

$$\mathbb{E}_{Q_{XY}}\left[\tfrac{P_X}{Q_X}\log\tfrac{P_X}{Q_X}\right] = \sum_x Q_X(x)\frac{P_X(x)}{Q_X(x)}\log\frac{P_X(x)}{Q_X(x)}\underbrace{\sum_y K_{Y|X=x}(y)}_{=1} = D\big(P_X \,\|\, Q_X\big). \qquad \square$$

Remark 4.7. Processing can only destroy distinguishability: no transformation of the data — deterministic or randomised — can make two hypotheses easier to tell apart than they were at the source. This is the fact that makes the converse of Stein's lemma believable before it is proved.

5 Hypothesis testing

Q. *State hypothesis testing for H_0 vs H_1 and define the two types of error.*

Definition 5.1 (Testing setup)**.** Let $X^n \overset{\text{iid}}{\sim} P$ and consider

$$H_0 :\ P = P_0 \quad \text{(null)}, \qquad H_1 :\ P = P_1 \quad \text{(alternative)}.$$

A *test* is a function

$$\varphi(x_1,\dots,x_n) = \begin{cases} 0, & \text{you believe } H_0 \text{ is true,} \\ 1, & \text{you believe } H_1 \text{ is true.} \end{cases}$$

The two error probabilities are

$$\underbrace{\alpha_n(\varphi) := P_0\big(\varphi(X^n)=1\big)}_{\text{Type I (false alarm)}}, \qquad \underbrace{\beta_n(\varphi) := P_1\big(\varphi(X^n)=0\big)}_{\text{Type II (missed detection)}}.$$

Definition 5.2 (Neyman–Pearson formulation)**.** Fix $\varepsilon > 0$ and restrict to tests with $\alpha_n(\varphi) \le \varepsilon$. Define

$$\beta_n^*(\varepsilon) := \min_{\varphi\,:\,\alpha_n(\varphi)\le\varepsilon} \beta_n(\varphi).$$

(The minimum is attained: on a finite alphabet there are finitely many tests on $\mathcal{X}^n$.)

Remark 5.3. Informally: if $\varepsilon \to 0$ as $n \to \infty$ then β_n may be driven to 1 or to 0 depending on how fast; if ε is held *fixed*, then $\beta_n^*(\varepsilon) \approx 2^{-nD\left(P_0 \,\|\, P_1\right)} \to 0$. Making this precise is Stein's lemma (Section 7).

Remark 5.4 (Lagrangian view)**.** One way to solve the constrained problem is $\mathcal{L}(\varphi,\lambda) = \beta_n(\varphi) + \lambda\big(\alpha_n(\varphi) - \varepsilon\big)$. By complementary slackness, $\lambda > 0 \Rightarrow \alpha_n(\varphi) = \varepsilon$: at the optimum the Type-I budget is spent in full.

5.1 The Neyman–Pearson lemma

Q. *State and prove the Neyman–Pearson lemma.* (Hint: $\alpha_n(\tau) = \mathbb{E}_{P_0}[\mathbf{1}(\varphi_\tau = 1)] = \mathbb{E}_{P_0}\big[\mathbf{1}\big(\frac{P_0}{P_1} \le \tau\big)\big] \le \tau\,\mathbb{E}_{P_1}\big[\mathbf{1}\big(\frac{P_0}{P_1} \le \tau\big)\big] = \tau\big(1-\beta_n(\tau)\big)$.)

Theorem 5.5 (Neyman–Pearson)**.** *Likelihood ratio tests are Pareto optimal. Fix* $0 < \tau < +\infty$ *and let*

$$\varphi_\tau(x^n) = \begin{cases} 0, & \dfrac{P_0(x^n)}{P_1(x^n)} > \tau, \\ 1, & \dfrac{P_0(x^n)}{P_1(x^n)} \le \tau, \end{cases} \qquad \textit{i.e.} \qquad \varphi_\tau = \mathbf{1}\Big(\tfrac{P_0(x^n)}{P_1(x^n)} \le \tau\Big).$$

Then for every test $\tilde\varphi$ *with* $\alpha_n(\tilde\varphi) < \alpha_n(\varphi_\tau)$ *we have* $\beta_n(\tilde\varphi) > \beta_n(\varphi_\tau)$.

Proof. Partition the discrepancy between the two tests:

$$\mathbf{1}_{\{\varphi_\tau=1\}} - \mathbf{1}_{\{\tilde\varphi=1\}} = \begin{cases} 0, & \varphi_\tau = \tilde\varphi, \\ 1, & \varphi_\tau = 1,\ \tilde\varphi = 0 \quad (\text{call this set } \mathcal{A}), \\ -1, & \varphi_\tau = 0,\ \tilde\varphi = 1 \quad (\text{call this set } \mathcal{B}). \end{cases}$$

Then

$$0 < \alpha_n(\varphi_\tau) - \alpha_n(\tilde\varphi) = \mathbb{E}_{P_0}\big[\mathbf{1}_{\mathcal{A}} - \mathbf{1}_{\mathcal{B}}\big] = \sum_{x^n\in\mathcal{A}} P_0(x^n) - \sum_{x^n\in\mathcal{B}} P_0(x^n).$$

On $\mathcal{A}$ we have $\varphi_\tau = 1$, i.e. $P_0(x^n) \le \tau P_1(x^n)$; on $\mathcal{B}$ we have $\varphi_\tau = 0$, i.e. $P_0(x^n) > \tau P_1(x^n)$. Hence

$$\sum_{\mathcal{A}} P_0 - \sum_{\mathcal{B}} P_0 \;\le\; \tau\Big(\sum_{\mathcal{A}} P_1 - \sum_{\mathcal{B}} P_1\Big) = \tau\,\mathbb{E}_{P_1}\big[\mathbf{1}_{\mathcal{A}} - \mathbf{1}_{\mathcal{B}}\big] = \tau\Big(P_1(\varphi_\tau = 1) - P_1(\tilde\varphi = 1)\Big).$$

Since $P_1(\varphi = 1) = 1 - \beta_n(\varphi)$, the bracket equals $\big(1-\beta_n(\varphi_\tau)\big) - \big(1-\beta_n(\tilde\varphi)\big) = \beta_n(\tilde\varphi) - \beta_n(\varphi_\tau)$. Therefore

$$0 < \tau\big(\beta_n(\tilde\varphi) - \beta_n(\varphi_\tau)\big),$$

and since $\tau > 0$ we conclude $\beta_n(\tilde\varphi) > \beta_n(\varphi_\tau)$. □

Remark 5.6. The hint in the margin is the same computation in compressed form: on the acceptance-of-H_1 region, $P_0 \le \tau P_1$ pointwise, so $\alpha_n(\tau) \le \tau\big(1-\beta_n(\tau)\big)$. Reducing τ trades Type-I error against Type-II error along the optimal frontier.

6 Divergence typical set

Definition 6.1 (Divergence typical set)**.** Fix $\varepsilon > 0$ and $n \in \mathbb{N}$. Put

$$T_\varepsilon^n(P_0, P_1) := \Big\{x^n \in \mathcal{X}^n \;:\; \Big|\tfrac{1}{n}\log\tfrac{P_0(x^n)}{P_1(x^n)} - D\big(P_0 \,\big\|\, P_1\big)\Big| < \varepsilon\Big\}.$$

This is "typical" because under P_0 the log-likelihood ratio is a sample average of i.i.d. terms with mean $D\big(P_0 \,\|\, P_1\big)$:

$$\frac{1}{n}\log\frac{P_0(X^n)}{P_1(X^n)} = \frac{1}{n}\sum_{i=1}^n \log\frac{P_0(X_i)}{P_1(X_i)} \;\xrightarrow{\;\mathbb{P}\;}\; D\big(P_0 \,\big\|\, P_1\big) \qquad \text{under } P_0.$$

Proposition 6.2 (Properties)**.** *Assume $P_0 \ll P_1$ and $D\big(P_0 \,\big\|\, P_1\big) < \infty$, and write $D = D\big(P_0 \,\big\|\, P_1\big)$.*

1) For all n and all $x^n \in T_\varepsilon^n(P_0, P_1)$,

$$P_0(x^n)\, 2^{-n(D+\varepsilon)} \;\le\; P_1(x^n) \;\le\; P_0(x^n)\, 2^{-n(D-\varepsilon)}.$$

2) For every $\varepsilon > 0$ and every $\delta > 0$, for n large enough, $P_0\big(T_\varepsilon^n(P_0, P_1)\big) > 1 - \delta$.

3) For all $n \in \mathbb{N}$, $P_1\big(T_\varepsilon^n(P_0, P_1)\big) \le 2^{-n(D-\varepsilon)}$.

4) For every $\varepsilon > 0$ and $\delta > 0$, for n large enough, $P_1\big(T_\varepsilon^n(P_0, P_1)\big) > (1-\delta)\, 2^{-n(D+\varepsilon)}$.

Proof. **1)** Rearranging the definition: $x^n \in T_\varepsilon^n$ means $D - \varepsilon < \frac{1}{n} \log \frac{P_0(x^n)}{P_1(x^n)} < D + \varepsilon$, i.e. $2^{n(D-\varepsilon)} < \frac{P_0(x^n)}{P_1(x^n)} < 2^{n(D+\varepsilon)}$.

2) By the WLLN applied to the i.i.d. summands $\log \frac{P_0(X_i)}{P_1(X_i)}$, which have mean D: for all $\varepsilon, \delta > 0$ there is $N_{\varepsilon,\delta}$ such that for all $n \ge N_{\varepsilon,\delta}$,

$$P_0\Big(\Big|\tfrac{1}{n} \log \tfrac{P_0(X^n)}{P_1(X^n)} - D\Big| < \varepsilon\Big) > 1 - \delta.$$

3) By 1),

$$P_1\big(T_\varepsilon^n\big) = \sum_{x^n \in T_\varepsilon^n} P_1(x^n) \;\le\; \sum_{x^n \in T_\varepsilon^n} P_0(x^n)\, 2^{-n(D-\varepsilon)} = 2^{-n(D-\varepsilon)} \underbrace{P_0\big(T_\varepsilon^n\big)}_{\le 1} \;\le\; 2^{-n(D-\varepsilon)}.$$

4) By 1) and then 2),

$$P_1\big(T_\varepsilon^n\big) = \sum_{x^n \in T_\varepsilon^n} P_1(x^n) \;\ge\; 2^{-n(D+\varepsilon)} P_0\big(T_\varepsilon^n\big) \;>\; (1-\delta)\, 2^{-n(D+\varepsilon)}$$

for n large enough. □

7 Stein's lemma

Q. *State and prove Stein's lemma.* (Hint: use the divergence typical set.)

Theorem 7.1 (Stein's lemma)**.** *Let $P_0 \ll P_1$ with $D := D\big(P_0 \,\big\|\, P_1\big) < \infty$, and fix $\varepsilon \in (0,1)$. Then*

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_n^*(\varepsilon) = D\big(P_0 \,\big\|\, P_1\big).$$

Equivalently, for large n, $\beta_n^*(\varepsilon) = 2^{-n(D+o(1))}$.

7.1 Achievability (upper bound on β_n^*)

Achievability. Fix any $\varepsilon' > 0$ and take the typical-set test

$$\varphi_{\varepsilon'}(x^n) = \begin{cases} 0, & x^n \in T_{\varepsilon'}^n(P_0, P_1), \\ 1, & \text{else.} \end{cases}$$

Type I. By Proposition 6.2.2 (with the deficit parameter set to ε), for n large enough,

$$\alpha_n(\varphi_{\varepsilon'}) = P_0\Big(\big(T^n_{\varepsilon'}\big)^c\Big) < \varepsilon,$$

so $\varphi_{\varepsilon'}$ is admissible in the definition of $\beta^*_n(\varepsilon)$.

Type II. By Proposition 6.2.3,

$$\beta_n(\varphi_{\varepsilon'}) = P_1\big(T^n_{\varepsilon'}\big) \le 2^{-n(D-\varepsilon')}.$$

Hence $\beta^*_n(\varepsilon) \le 2^{-n(D-\varepsilon')}$, i.e.

$$-\frac{1}{n}\log\beta^*_n(\varepsilon) \;\ge\; D-\varepsilon' \qquad \text{for all large } n,$$

so $\liminf_n -\frac{1}{n}\log\beta^*_n(\varepsilon) \ge D-\varepsilon'$. Since $\varepsilon'>0$ was arbitrary, letting $\varepsilon'\downarrow 0$ gives $\liminf_n -\frac{1}{n}\log\beta^*_n(\varepsilon) \ge D$. □

7.2 Optimality (lower bound on β^*_n)

Optimality. Fix any $\varepsilon'>0$ and let $\tilde\varphi$ be *any* test with $\alpha_n(\tilde\varphi)\le\varepsilon$. Write $T := T^n_{\varepsilon'}(P_0,P_1)$ and let $\hat\varphi_{\varepsilon'}$ be the typical-set test of the previous subsection, so that $\{\hat\varphi_{\varepsilon'}=0\} = \{X^n\in T\}$.

Step 1: restrict to the typical set and change measure.

$$\begin{aligned}
\beta_n(\tilde\varphi) = P_1\big(\tilde\varphi=0\big) \;&\ge\; P_1\big(\tilde\varphi=0,\ \hat\varphi_{\varepsilon'}=0\big) = \sum_{\substack{x^n:\ \tilde\varphi(x^n)=0\\ x^n\in T}} P_1(x^n)\\
&\ge \sum_{\substack{x^n:\ \tilde\varphi(x^n)=0\\ x^n\in T}} 2^{-n(D+\varepsilon')}P_0(x^n) \qquad \text{(Proposition 6.2.1)}\\
&= 2^{-n(D+\varepsilon')}\,P_0\big(\tilde\varphi=0,\ \hat\varphi_{\varepsilon'}=0\big).
\end{aligned}$$

Step 2: the remaining probability is bounded below. Set $\eta := \frac{1-\varepsilon}{2} > 0$. By Proposition 6.2.2 (deficit η), for n large enough $P_0(T^c)\le\eta$. Then

$$\begin{aligned}
P_0\big(\tilde\varphi=0,\ \hat\varphi_{\varepsilon'}=0\big) &= 1 - P_0\big(\tilde\varphi=1 \ \text{ or } \ \hat\varphi_{\varepsilon'}=1\big)\\
&\ge 1 - P_0\big(\tilde\varphi=1\big) - P_0\big(\hat\varphi_{\varepsilon'}=1\big) && \text{(union bound)}\\
&\ge 1-\varepsilon - P_0\big(\hat\varphi_{\varepsilon'}=1\big) && (\alpha_n(\tilde\varphi)\le\varepsilon)\\
&= 1-\varepsilon-P_0\big(T^c\big) \;\ge\; 1-\varepsilon-\eta = \frac{1-\varepsilon}{2} \;>\; 0.
\end{aligned}$$

Step 3: conclude. Combining,

$$\beta_n(\tilde\varphi) \;\ge\; \frac{1-\varepsilon}{2}\,2^{-n(D+\varepsilon')}$$

for all large n, uniformly over admissible $\tilde\varphi$; hence the same bound holds for $\beta^*_n(\varepsilon)$. Taking logarithms,

$$-\frac{1}{n}\log\beta^*_n(\varepsilon) \;\le\; D+\varepsilon' - \frac{1}{n}\log\frac{1-\varepsilon}{2} \;\xrightarrow[n\to\infty]{}\; D+\varepsilon',$$

so $\limsup_n -\frac{1}{n}\log\beta^*_n(\varepsilon)\le D+\varepsilon'$. Since $\varepsilon'>0$ was arbitrary, $\limsup_n -\frac{1}{n}\log\beta^*_n(\varepsilon)\le D$.

Together with the achievability bound this proves Theorem 7.1. □

Remark 7.2 (Interpretation)**.** Stein's lemma says the best achievable Type-II error decays exponentially at rate exactly $D(P_0 \,\|\, P_1)$, and — remarkably — this rate does not depend on the Type-I budget ε, as long as $\varepsilon \in (0, 1)$ is held fixed. **The divergence is thus the operational measure of how distinguishable two hypotheses are**, which is also why the data processing inequality (Proposition 4.6) is the right structural property for it to have.