4 Application to statistics: Binary hypothesis testing

Intended learning outcomes:

- You can compute the minimal error probability in binary hypothesis testing with known priors, and understand its relationship with the total variation distance.
- You are able to apply and analyse threshold tests.
- You understand the setup of symmetric and asymmetric binary hypothesis testing.
- You can apply the Chernoff exponent and Stein's lemma.

Book reference: Chapter 11, Sections 11.7–11.9 in Cover & Thomas[1], but we are not following it too closely.

[1] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991. ISBN 9780471748823. DOI: 10.1002/047174882X

4.1 Binary hypothesis testing

We consider binary hypothesis testing where we try to distinguish between two models of a random process. The random process produces a sequence of random variables $\boldsymbol{X} = (X_1, X_2, \ldots)$ that are independently drawn from some (unknown) probability distribution $Q \in \mathcal{P}(\mathcal{X})$, where we take $\mathcal{X}$ to be any discrete set. Consider the hypothesis test between two hypotheses, H_0 and H_1:

$$\begin{aligned} H_0 &: Q = P_0 \\ H_1 &: Q = P_1, \end{aligned} \tag{4.1}$$

where $P_0, P_1 \in \mathcal{P}(\mathcal{X})$ are two candidate probability distributions (or models) of the process creating the random outcome. Our goal is to deduce, from the observation of the random sequence $\boldsymbol{X}$, which of the two hypothesis is correct. H_0 is usually called the *null-hypothesis* and H_1 the *alternate hypothesis*.

A (deterministic) *test* for the sequence $X^{(n)} = (X_1, X_2, \ldots, X_n)$ is a region $\mathcal{A}_n \subset \mathcal{X}^n$. We say that the alternate hypothesis is *accepted* for this test if the observed sequence satisfies $(x_1, x_2, \ldots, x_n) \in \mathcal{A}_n$, and it is *rejected* otherwise. If the alternate hypothesis is rejected the null-hypothesis is maintained. We can then define two kinds of errors:

$$\alpha_n(\mathcal{A}_n) := P_0^n(\mathcal{A}_n) \tag{4.2}$$

$$\beta_n(\mathcal{A}_n) := 1 - P_1^n(\mathcal{A}_n) \tag{4.3}$$

The *error of the first kind* or *type-I error*, $\alpha_n(\mathcal{A}_n)$, captures the acceptance of the alternate hypothesis even if the null-hypothesis is true. The *error of the second kind* or *type-2 error*, $\beta_n(\mathcal{A}_n)$, captures the rejection of the alternate hypothesis even though it is true.

Ideally we would like to devise a sequence of tests such that both of these errors are small, and get smaller as n increases. We can compute the optimal average error assuming an uniform (or unbiased) prior on the two distributions.

Can you formulate this problem in the general framework of probability theory as covered in Chapter 0? What is a test in this framework?

For two pmfs P_0 and P_1 and $n \in \mathbb{N}$, we define the *optimal unbiased average error* as

$$\epsilon^*_{\mathrm{sym},n}(P_0, P_1) := \frac{1}{2} \min_{\mathcal{A}_n \subset \mathcal{X}^n} \Big(\alpha_n(\mathcal{A}_n) + \beta_n(\mathcal{A}_n) \Big) . \tag{4.4}$$

Here uniform prior means that the probability we assign to the two hypotheses prior to observing the random sequence is equal, and thus $\epsilon^*_{\mathrm{sym},n}$ is indeed the probability of making a wrong decision. However, as their names indicates, often these two hypotheses are not treated on the same footing. Indeed, the question can be easily generalised to the case when the prior over the two hypothesis is not uniform. If $p \in (0,1)$ is the probability that H_0 is correct, we define

$$\epsilon^*_{p,n}(P_0, P_1) := \min_{\mathcal{A}_n \subset \mathcal{X}^n} \Big(p\alpha_n(\mathcal{A}_n) + (1-p)\beta_n(\mathcal{A}_n) \Big) . \tag{4.5}$$

The probability p is called the prior in this setting. For such cases it is natural to look at a somewhat different and inherently asymmetric formulation of the problem. Namely, we simply require that one of the errors (by convention the error of the first kind) is upper bounded by a constant ϵ and ask how small we can make the other error.

Example. *Assume the alternate hypothesis is that a patient is suffering from COVID-19, and the null-hypothesis is that this is not the case. The error of the first kind is then a false positive and the error of the second kind is a false negative. If we devise a test distinguishing these two hypothesis we are probably more tolerant of false positives than false negatives.*

For two pmfs P_0 and P_1 and $n \in \mathbb{N}$ and $\epsilon \in (0,1)$, we define

$$\beta^*_n(\epsilon; P_0, P_1) := \min\{\beta_n(\mathcal{A}_n) : \alpha_n(\mathcal{A}_n) \le \epsilon\}, \tag{4.6}$$

where $\mathcal{A}_n$ runs through all subsets of $\mathcal{X}^n$.

Asymmetric hypothesis testing also allows us to deal with the situation when we do not know the prior probabilities. In that case the sum (or probabilistic mixture) of the two errors does not make sense and we need to look at the errors independently. We can, however, still ask the question how these two errors trade off against each other. This is done by analysing $\beta_n^*(\epsilon)$, and in particular by looking at its asymptotics for large n.

4.2 Optimal tests for the one-shot setting

We will first try to understand the one-shot setting, i.e. we set $n = 1$. Notably, this actually also covers the cases where $n > 1$ in the sense that we can always see the joint distributions P_0^n and P_1^n as our two hypothesis. However, in the one-shot setting we do not have any i.i.d. structure to work with, which often allows us to simplify the problem when n is large.

It turns out that the minimal error probability $\epsilon^*_{\mathrm{sym},1}$ is related to the total variation distance, and we will explore this relation first. We can then deduce a much more general property of optimal tests, called the Neymann–Pearson lemma.

4.2.1 Hypothesis testing and total variation distance

In the following will see that $\epsilon^*_{\mathrm{sym},1}$ can be expressed in terms of the *total variation distance* (tvd) between two pmfs P_0 and P_1. Recall that the total variational distance is closely related to the 1-norm, i.e., $\delta_{\mathrm{tvd}}(P_0, P_1) = \frac{1}{2}\|P_0 - P_1\|_1$. We can now state the following result for binary hypothesis testing with general priors.

Proposition 4.1. *The one-shot optimal average error is given by*

$$\epsilon^*_{p,1}(P_0, P_1) = \frac{1}{2}\Big(1 - \big\|pP_0 - (1-p)P_1\big\|_1\Big)\,. \tag{4.7}$$

In particular, if $p = 1 - p = \frac{1}{2}$, we have $\epsilon^*_{\mathrm{sym},1} = \frac{1}{2}\left(1 - \delta_{\mathrm{tvd}}(P_0, P_1)\right)$.

This gives a clear operational interpretation for the total variation distance, which is a widely used distance measure in statistics. On the one hand, when $P_0 = P_1$ the total variation distance vanishes and the best thing we can do is a random guess. On the other hand, when P_0 and P_1 are orthogonal, then we can distinguish them perfectly and the error vanishes.

Proof of Proposition 4.1. First we observe the following relations:

$$\begin{aligned}
\epsilon^*_{p,1} &= \min_{\mathcal{A}\subset\mathcal{X}} \left(pP_0(\mathcal{A}) + (1-p)P_1(\mathcal{A}^c)\right) && (4.8)\\
&= p - \max_{\mathcal{A}\subset\mathcal{X}} \left(pP_0(\mathcal{A}^c) - (1-p)P_1(\mathcal{A}^c)\right) && (4.9)\\
&= 1 - p - \max_{\mathcal{A}\subset\mathcal{X}} \left((1-p)P_1(\mathcal{A}) - pP_0(\mathcal{A})\right) && (4.10)
\end{aligned}$$

Combining the two relations we get

$$2\epsilon^*_{p,1} = 1 - \max_{\mathcal{A}\subset\mathcal{X}} \big((1-p)P_1(\mathcal{A}) - pP_0(\mathcal{A})\big) - \max_{\mathcal{A}\subset\mathcal{X}} \big(pP_0(\mathcal{A}^c) - (1-p)P_1(\mathcal{A}^c)\big) \tag{4.11}$$

At this point we can determine which sets achieve the maximum in these two optimisation. Clearly both maxima are achieved by the set

$$\mathcal{A}_* = \{x \in \mathcal{X} : (1-p)P_1(x) \geq pP_0(x)\}\,. \tag{4.12}$$

The above expression then simplifies to

$$2\epsilon^*_{p,1} = 1 - \sum_{x\in\mathcal{A}_*} (1-p)P_1(x) - pP_0(x) - \sum_{x\in\mathcal{A}^c_*} pP_0(x) - (1-p)P_1(x) \tag{4.13}$$

$$= 1 - \sum_{x\in\mathcal{X}} \big|pP_0(x) - (1-p)P_1(x)\big| \tag{4.14}$$

$$= 1 - \big\|pP_0 - (1-p)P_1\big\|_1\,. \tag{4.15}$$

Dividing both sides by 2 then yields the desired result. □

4.2.2 *The Neyman-Pearson lemma*

The test in Eq. (4.12) is of the form

$$\mathcal{A}_* = \left\{x \in \mathcal{X} : \log\frac{P_0(x)}{P_1(x)} \leq T\right\}\,. \tag{4.16}$$

where $T = \log\frac{1-p}{p}$ is a threshold and $\log\frac{P_0(x)}{P_1(x)}$ is the well-known log-likelihood ratio. We can show that optimal tests must always have this form.

Example. *Consider a DMS with either $P_1 = (\frac{1}{2}, \frac{1}{2})$ or $P_2 = (\frac{3}{4}, \frac{1}{4})$. For $n = 2$ we use the shorthand notation $00, 01, 10, 11$ to denote the different possible sequences in $\mathcal{X}^2$. The possible threshold tests are:*

A	α_2	β_2	T
$\emptyset$	0	1	$(-\infty, 2\log\frac{2}{3})$
$\{00\}$	$\frac{1}{4}$	$\frac{7}{16}$	$[2\log\frac{2}{3}, \log\frac{4}{3})$
$\{00, 01, 10\}$	$\frac{3}{4}$	$\frac{1}{16}$	$[\log\frac{4}{3}, 2)$
$\mathcal{X}^2$	1	0	$[2, +\infty)$

Here $\alpha = \alpha(A)$ and $\beta = \beta(A)$ are two two kinds of errors and we give the range of T which produces this test.

Lemma 4.2 (Neymann-Pearson Lemma). *Let P_0, P_1 be two pmfs. For any $T \in \mathbb{R}$, define the region*

$$\mathcal{A}_*(T) = \left\{x : \log\frac{P_0(x)}{P_1(x)} \leq T\right\}\,. \tag{4.17}$$

Then, for any test $\mathcal{A}$ and any $T \in \mathbb{R}$, the following holds:

$$\alpha_1(\mathcal{A}) < \alpha_1(\mathcal{A}_*(T)) \implies \beta_1(\mathcal{A}) > \beta_1(\mathcal{A}_*(T)) \tag{4.18}$$

$$\beta_1(\mathcal{A}) < \beta_1(\mathcal{A}_*(T)) \implies \alpha_1(\mathcal{A}) > \alpha_1(\mathcal{A}_*(T)) \tag{4.19}$$

Proof. We fix T and write $\mathcal{A}_*(T) = \mathcal{A}_*$. For every $x \in \mathcal{X}$, we have

$$(\mathbf{1}\{x \in \mathcal{A}_*\} - \mathbf{1}\{x \in \mathcal{A}\})\,(P_0(x) - 2^T P_1(x)) \leq 0 \tag{4.20}$$

since $\mathbf{1}\{x \in \mathcal{A}_*\}$ indicates that $P_0(x) - 2^T P_1(x)$ is negative by the definition of $\mathcal{A}_*$. Now summing over $x \in \mathcal{X}$ we find

$$\alpha_1(\mathcal{A}_*) - \alpha_1(\mathcal{A}) - 2^T\,(1 - \beta_1(\mathcal{A}_*) - (1 - \beta_1(\mathcal{A}))) \leq 0\,, \tag{4.21}$$

or, equivalently, $\alpha_1(\mathcal{A}_*) - \alpha_1(\mathcal{A}) \leq 2^T(\beta_1(\mathcal{A}) - \beta_1(\mathcal{A}_*))$. Since $2^T > 0$ the conditions in Eq. (4.18) follows. □

4.3 *The Chernoff exponent in symmetric hypothesis testing*

When we look at n i.i.d. copies of the sample, distributed according to P_0^n or P_1^n, respectively, we make the at first sight surprising observation that these two distributions get closer and closer to orthogonal as $n \to \infty$ (unless $P_0 = P_1$, of course). Or, in other words, the total variation distance between P_0^n and P_1^n converges to 1 as $n \to \infty$.

We are interested how fast this convergence to zero is. We will show the following bound:

Proposition 4.3. *For any two pmfs P_0 and P_1, we have*

$$\lim_{n\to\infty} -\frac{1}{n}\log \epsilon^*_{\mathrm{sym},n}(P_0,P_1) \geq C(P_0,P_1)\,, \tag{4.22}$$

where we introduced the Chernoff distance *or Chernoff exponent,*

$$C(P_0,P_1) := -\min_{0\leq\lambda\leq 1}\log\sum_{x\in\mathcal{X}} P_0(x)^\lambda P_1(x)^{1-\lambda}\,. \tag{4.23}$$

One obviously needs to verify that $C(P_0,P_1)$ is positive. To do so, first note that it suffices to verify that

$$\sum_x \sqrt{P_0(x)P_1(x)} \leq 1$$

holds. This is however ensured by the Cauchy-Schwarz inequality (Prop. 0.8). Moreover, this ensures that $C(P_0,P_1) \geq 0$ with equality iff $P_0 = P_1$.

Note that this actually corresponds to an asymptotic upper bound on the probability of error, so it says that there exists a sequence of tests for which the error drops as $2^{-nC(P_0,P_1)}$. It turns out (but we will not show this here) that this is optimal, i.e., that equality holds in Eq. (4.22).

Proof of Proposition 4.3. To show the upper bound on the error probability, we first argue that, for every $\lambda \in [0,1]$:

$$\begin{aligned}
2\epsilon^*_{\mathrm{sym},n} &= \min_{\mathcal{A}_n\subset\mathcal{X}^n} P_0^n(\mathcal{A}_n) + P_1^n(\mathcal{A}_n^c) &(4.24)\\
&= \sum_{x^n\in\mathcal{X}^n}\min\{P_0^n(x^n), P_1^n(x^n)\} &(4.25)\\
&\leq \sum_{x^n\in\mathcal{X}^n} P_0^n(x^n)^\lambda P_1^n(x^n)^{1-\lambda} &(4.26)\\
&= \sum_{x_1\in\mathcal{X}}\dots\sum_{x_n\in\mathcal{X}} P_0(x_1)^\lambda P_1(x_1)^{1-\lambda}\dots P_0(x_n)^\lambda P_1(x_n)^{1-\lambda} &(4.27)\\
&= \left(\sum_{x\in\mathcal{X}} P_0(x)^\lambda P_1(x)^{1-\lambda}\right)^n &(4.28)
\end{aligned}$$

We take the logarithm on both sides and divide through n to get

$$\frac{1}{n}\log\epsilon^*_{\mathrm{sym},n} \leq \log\sum_{x\in\mathcal{X}} P_0(x)^\lambda P_1(x)^{1-\lambda} - \frac{1}{n} \tag{4.29}$$

What changes when we do the same analysis for $\epsilon^*_{p,n}$?

The last term vanishes in the limit $n \to \infty$, and thus the bound (4.22) follows by maximising the right-hand side over all $\lambda \in [0,1]$. □

4.4 Stein's lemma in asymmetric hypothesis testing

For simplicity we assume that $D(P_0 \| P_1) < \infty$ in the following, as otherwise by definition of the relative entropy there are some $x \in \mathcal{X}$ with $P_0(x) > 0$ but $P_1(x) = 0$, and, as we will see in the homework, it is possible to come up with tests that have $\beta_n^*(\epsilon) = 0$ for large enough n.

Under this assumption, our goal is to show that regardless of the constant upper bound ϵ on the type-I error, the type-II error behaves as

$$\beta_n^*(\epsilon) \approx 2^{-nD(P_0 \| P_1)}, \tag{4.30}$$

where the approximation is up to factors that are sub-exponential in n. This means that the optimal exponential rate at which the type-II error approaches zero is determined by the relative entropy (in first order), thus giving the relative entropy $D(P_0 \| P_1)$ a clear operational interpretation in statistics. Let us restate this as a theorem:

Theorem 4.4 (Chernoff-Stein Lemma)**.** *For every $\epsilon \in (0, 1)$,*

$$\lim_{n \to \infty} -\frac{1}{n} \log \beta_n^*(\epsilon) = D(P_0 \| P_1) \tag{4.31}$$

For the proof we will use the information spectrum method (see[2] for more information on that technique). Consider a random variable X that takes values on $\mathcal{X}$ and two pmfs $P_0, P_1 \in \mathcal{P}(\mathcal{X})$ as above. Recall that the log-likelihood ratio for the two pmfs is the random variable

$$Z = \log \frac{P_0(X)}{P_1(X)}, \tag{4.32}$$

where X (and thus Z) is distributed according to P_0. The log-likelihood ratio is an important random variable in the analysis of many different information processing tasks. We now introduce the following quantity:

For two pmfs P_0, P_1 and $\epsilon \in (0, 1)$ we define the *information-spectrum divergence* as

$$D_s^{\epsilon}(P_0 \| P_1) := \sup \{R \in \mathbb{R} : P_0[Z \le R] \le \epsilon\} \tag{4.33}$$

This quantity looks complicated at first sight, but it simply evaluates exactly where (the value R) we need to cut off the pmf for the *log-likelihood ratio*, Z, so that the probability that $Z \le R$ is at most ϵ. One could also simply see it as an inverse of the cumulative distribution function of Z. (This result can also be interpreted as a consequence of the Neyman-Pearson lemma, which states that all tests optimising the two types of errors are threshold tests for the log-likelihood ratio.)

[2] T. S. Han. *Information-Spectrum Methods in Information Theory*. Applications of Mathematics. Springer, 2002

Verify that the expectation value of Z under P_0 is $D(P_0 \| P_1)$.

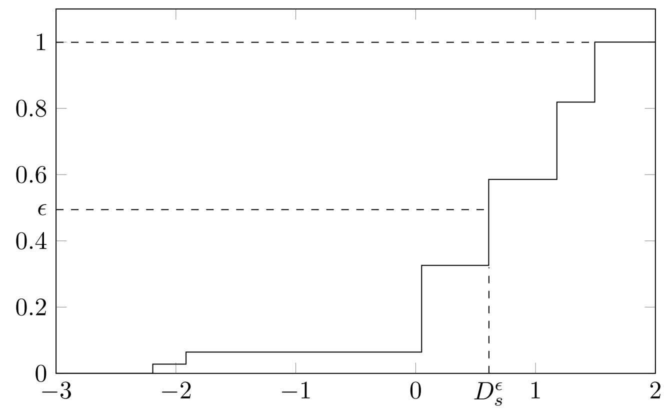

Figure 4.1: Example of the information spectrum. The plot shows the cumulative distribution of $Z = \log \frac{P(X)}{Q(X)}$ and the value of $D_s^\epsilon(P\|Q)$ for some example distributions.

Lemma 4.5. *Let $n \in \mathbb{N}$, $\epsilon \in (0,1)$ and $\delta \in (0, 1-\epsilon)$. The following two inequalities hold:*

$$D_s^\epsilon(P_0^n \| P_1^n) \le -\log \beta_n^*(\epsilon) \le D_s^{\epsilon+\delta}(P_0^n \| P_1^n) + \log \frac{1}{\delta}\,. \tag{4.34}$$

For $n = 1$ this gives bounds on asymmetric hypothesis testing for any two distributions P_0 and P_1, without using the i.i.d. structure. If one plugs in n-fold i.i.d. distributions instead this recovers the result for general n. This is an example of a *one-shot bound*, a generic bound on an information-theoretic quantity that can then be easily statistically analysed by taking advantage of an i.i.d. or similar structure.

Proof. To get the lower bound, we use a threshold test in the sense of Neymann-Pearson (see Lemma 4.2) of the form

$$\mathcal{A}_{T,n} := \{x^n \in \mathcal{X}^n : P_0^n(x^n) \le 2^T P_1^n(x^n)\}\,. \tag{4.35}$$

Let us choose $T = D_s^\epsilon(P_0^n \| P_1^n) - \mu$ for some $\mu > 0$ that can be chosen arbitrarily small. The reason we need this small slack $\mu > 0$ is simply that by definition of the supremum in (4.33) this ensures that we have $\alpha_n(\mathcal{A}_{R,n}) = P_0^n(\mathcal{A}_{R,n}) \le \epsilon$ for any $\mu > 0$, while the same might not necessarily be true for $\mu = 0$.[3] Moreover, we have

$$\begin{aligned}
\beta_n(\mathcal{A}_{T,n}) &= P_1^n(\mathcal{A}_{T,n}^c) && (4.36)\\
&= \sum_{x^n \in \mathcal{X}^n} P_1^n(x^n)\,\mathbf{1}\{P_0^n(x^n) > 2^T P_1^n(x^n)\} && (4.37)\\
&\le 2^{-T} \sum_{x^n \in \mathcal{X}^n} P_0^n(x^n)\,\mathbf{1}\{P_0^n(x^n) > 2^T P_1^n(x^n)\} && (4.38)\\
&\le 2^{-T}\,. && (4.39)
\end{aligned}$$

[3] Recall the definition of the supremum on sets that are not closed and note that the set of all $R \in \mathbb{R}$ satisfying $P[Z \le R] \le \epsilon$ is not closed in general.

This directly implies that $\beta_n^*(\epsilon) \leq 2^{-T}$, or, equivalently,

$$-\log \beta_n^*(\epsilon) \geq D_s^\epsilon(P_0^n \| P_1^n) - \mu\,. \tag{4.40}$$

Since this holds for all $\mu > 0$ we get the desired inequality.

To get the upper bound, let $\mathcal{A}_n$ be the optimal test for $\beta_n^*(\epsilon)$, i.e. we have $\alpha_n(\mathcal{A}_{T,n}) \leq \epsilon$ and $\beta_n(\mathcal{A}_{T,n}) = \beta_n^*(\epsilon)$. Note that due to Lemma 4.2 we can assume that the optimal test is a threshold test, but T is still to be determined. We can establish the following sequence of inequalities:

$$\begin{aligned}
&1 - P_0^n\left(\log \frac{P_0^n(X^n)}{P_1^n(X^n)} \leq R\right) \\
&\quad = \sum_{x^n \in \mathcal{X}^n} P_0^n(x^n)\, \mathbf{1}\{P_0^n(x^n) > 2^R P_1^n(x^n)\} && (4.41)\\
&\quad \geq \sum_{x^n \in \mathcal{X}^n} \left(P_0^n(x^n) - 2^R P_1^n(x^n)\right) \mathbf{1}\{P_0^n(x^n) > 2^R P_1^n(x^n)\} && (4.42)\\
&\quad \geq \sum_{x^n \in \mathcal{X}^n} \left(P_0^n(x^n) - 2^R P_1^n(x^n)\right) \mathbf{1}\{x^n \in \mathcal{A}_n^c\} && (4.43)\\
&\quad = P_0^n(\mathcal{A}_n^c) - 2^R P_1^n(\mathcal{A}_n^c) && (4.44)\\
&\quad = 1 - \alpha_n(\mathcal{A}_n) - 2^R \beta_n(\mathcal{A}_n) && (4.45)\\
&\quad \geq 1 - \epsilon - 2^R \beta_n^*(\epsilon)\,. && (4.46)
\end{aligned}$$

The critical step is to get from Eq. (4.42) to Eq. (4.43). To verify this, note that the test $1\{P_0^n(x^n) > 2^R P_1^n(x^n)\}$ is actually the one that maximises the sum since it cuts out all negative contributions. Any other test, including $\mathcal{A}_n^c$, can thus only reduce the sum.

Now, if we choose $R = \log \delta - \log \beta_n^*(\epsilon)$, the above implies that

$$P_0^n\,(Z \leq R) \leq \epsilon + \delta \tag{4.47}$$

and thus we have $D_s^{\epsilon+\delta}(P_0^n \| P_1^n) \geq R = -\log \beta_n^*(\epsilon) + \log \delta$, which is the desired upper bound. □

In the homework you have shown that the following asymptotic limit holds. The limit is essentially a direct consequence of the law of large numbers applied for the random variable $Z = \sum_{i=1}^n \log P_0(X_i) - \log P_1(X_i)$, where X_i are i.i.d. distributed according to the law P_0.

As an aside, we can evaluate the quantity on the left, $\frac{1}{n} D_s^\epsilon(P_0^n \| P_1^n)$, even to higher orders in n using the central limit theorem. While we will not need this here, analysing such higher order terms has been a fruitful area of research recently as it allows us to make more precise statements about optimal errors for smaller n, and thus for practical settings where we are far from the asymptotic setting of very large n.

Lemma 4.6. *Let $P_0, P_1 \in \mathcal{P}(X)$ such that $D(P_0 \| P_1) < \infty$ and $\epsilon \in (0,1)$. Then,*

$$\lim_{n \to \infty} \frac{1}{n} D_s^\epsilon(P_0^n \| P_1^n) = D(P_0 \| P_1)\,. \tag{4.48}$$

Proof of Theorem 4.4. The proof of the theorem is evident once we combine Lemma 4.5 and Lemma 4.6. Namely, from Lemma 4.5 we get

$$\frac{1}{n} D_s^{\epsilon}(P_0^n \| P_1^n) \leq -\frac{1}{n} \log \beta_n(\epsilon) \leq \frac{1}{n} D_s^{\epsilon+\delta}(P_0^n \| P_1^n) + \frac{1}{n} \log \frac{1}{\delta} \qquad (4.49)$$

and in the limit $n \to \infty$ both the lower and upper bound converge to the relative entropy by Lemma 4.6. $\square$