§ 13. Hypothesis testing asymptotics II

Setup:

$$\begin{aligned}&H_0 : X^n \sim P_{X^n} \qquad H_1 : X^n \sim Q_{X^n} \quad \text{(i.i.d.)}\\ &\text{test } P_{Z|X^n} : \mathcal{X}^n \to \{0,1\}\\ &\text{specification: } 1-\alpha = \pi_{1|0}^{(n)} \le 2^{-nE_0} \qquad \beta = \pi_{0|1}^{(n)} \le 2^{-nE_1}\end{aligned}$$

Bounds:

- achievability (Neyman Pearson)

$$\alpha = 1 - \pi_{1|0} = P_{X^n}[F_n > \tau], \qquad \beta = \pi_{0|1} = Q_{X^n}[F_n > \tau]$$

- converse (strong)

$$\forall (\alpha,\beta) \text{ achievable, } \alpha - \gamma\beta \le P_{X^n}[F > \log\gamma]$$

where

$$F = \log \frac{dP_{X^n}}{dQ_{X^n}}(X^n),$$

13.1 (E_0, E_1)-Tradeoff

Goal:

$$1-\alpha \le 2^{-nE_0}, \quad \beta \le 2^{-nE_1}.$$

Our goal in the Chernoff regime is to find the best tradeoff, which we formally define as follows (compare to Stein's exponent in Lecture 11)

$$\begin{aligned}E_1^*(E_0) &\triangleq \sup\{E_1 : \exists n_0, \forall n \ge n_0, \exists P_{Z|X^n} \text{ s.t. } \alpha > 1-2^{-nE_0}, \beta < 2^{-nE_1}, \}\\ &= \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_{1-2^{-nE_0}}(P^n, Q^n)}\end{aligned}$$

Define

$$T = \log\frac{dQ}{dP}(X), \quad T_k = \log\frac{dQ}{dP}(X_k), \quad \text{thus } \log\frac{dQ^n}{dP^n}(X^n) = \sum_{k=1}^n T_k$$

Log MGF of T under P (again assumed to be finite and also $T \ne \text{const}$ since $P \ne Q$):

$$\begin{aligned}\psi_P(\lambda) &= \log \mathbb{E}_P[e^{\lambda T}]\\ &= \log \sum_x P(x)^{1-\lambda} Q(x)^\lambda = \log \int (dP)^{1-\lambda}(dQ)^\lambda\\ \psi_P^*(\theta) &= \sup_{\lambda\in\mathbb{R}} \theta\lambda - \psi_P(\lambda)\end{aligned}$$

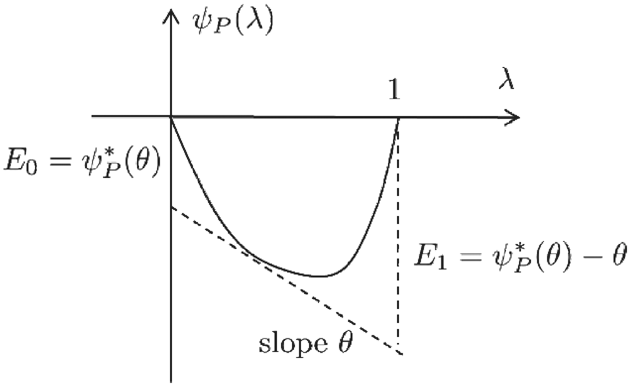
Note that since $\psi_P(0) = \psi_P(1) = 0$ from convexity $\psi_P(\lambda)$ is finite on $0 \le \lambda \le 1$. Furthermore, assuming $P \ll Q$ and $Q \ll P$ we also have that $\lambda \mapsto \psi_P(\lambda)$ continuous everywhere on $[0,1]$ (on $(0,1)$ it follows from convexity, but for boundary points we need more detailed arguments). Consequently, all the results in this section apply under just the conditions of $P \ll Q$ and $Q \ll P$. However, since in previous lecture we were assuming that log-MGF exists for all λ, we will only present proofs under this extra assumption.

Theorem 13.1. *Let $P \ll Q$, $Q \ll P$, then*

$$E_0(\theta) = \psi_P^*(\theta), \qquad E_1(\theta) = \psi_P^*(\theta) - \theta \tag{13.1}$$

parametrized by $-D(P\|Q) \le \theta \le D(Q\|P)$ characterizes the best exponents on the boundary of achievable (E_0, E_1).

Note: The geometric interpretation of the above theorem is shown in the following picture, which rely on the properties of $\psi_P(\lambda)$ and $\psi_P^*(\theta)$. Note that $\psi_P(0) = \psi_P(1) = 0$. Moreover, by Theorem 11.3 (Properties of ψ_X^*), $\theta \mapsto E_0(\theta)$ is increasing, $\theta \mapsto E_1(\theta)$ is decreasing.



Remark 13.1 (Rényi divergence)**.** Rényi defined a family of divergence indexed by $\lambda \ne 1$

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda - 1} \log \mathbb{E}_Q \left[\left(\frac{dP}{dQ}\right)^\lambda\right] \ge 0.$$

which generalizes Kullback-Leibler divergence since $D_\lambda(P\|Q) \xrightarrow{\lambda \to 1} D(P\|Q)$. Note that $\psi_P(\lambda) = (\lambda - 1) D_\lambda(Q\|P) = -\lambda D_{1-\lambda}(P\|Q)$. This provides another explanation that ψ_P is negative between 0 and 1, and the slope at endpoints is: $\psi_P'(0) = -D(P\|Q)$ and $\psi_P'(1) = D(Q\|P)$.

Corollary 13.1 (Bayesian criterion)**.** *Fix a prior (π_0, π_1) such that $\pi_0 + \pi_1 = 1$ and $0 < \pi_0 < 1$. Denote the optimal Bayesian (average) error probability by*

$$P_e^*(n) \triangleq \inf_{P_{Z|X^n}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}$$

with exponent

$$E \triangleq \lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P_e^*(n)}.$$

Then

$$E = \max_\theta \min(E_0(\theta), E_1(\theta)) = \psi_P^*(0) = -\inf_{\lambda \in \mathbb{R}} \psi_P(\lambda),$$

regardless of the prior, and $\psi_P^*(0) \triangleq C(P,Q)$ is called the Chernoff exponent.

Proof of Theorem 13.1. The idea is to apply the large deviation theory to iid sum $\sum_{k=1}^n T_k$. Specifically, let's rewrite the bounds in terms of T:

- Achievability (Neyman Pearson)

$$\text{let } \tau = -n\theta, \quad \pi_{1|0}^{(n)} = P\left[\sum_{k=1}^n T_k \geq n\theta\right] \quad \pi_{0|1}^{(n)} = Q\left[\sum_{k=1}^n T_k < n\theta\right]$$

- Converse (strong)

$$\text{let } \gamma = 2^{-n\theta}, \quad \pi_{1|0} + 2^{-n\theta}\pi_{0|1} \geq P\left[\sum_{k=1}^n T_k \geq n\theta\right]$$

Achievability: Using Neyman Pearson test, for fixed $\tau = -n\theta$, apply the large deviation theorem:

$$\begin{aligned} 1 - \alpha = \pi_{1|0}^{(n)} &= P\left[\sum_{k=1}^n T_k \geq n\theta\right] = 2^{-n\psi_P^*(\theta)+o(n)}, \qquad \text{for } \theta \geq \mathbb{E}_P T = -D(P\|Q) \\ \beta = \pi_{0|1}^{(n)} &= Q\left[\sum_{k=1}^n T_k < n\theta\right] = 2^{-n\psi_Q^*(\theta)+o(n)}, \qquad \text{for } \theta \leq \mathbb{E}_Q T = D(Q\|P) \end{aligned}$$

Notice that by the definition of T we have

$$\begin{aligned} &\psi_Q(\lambda) = \log \mathbb{E}_Q[e^{\lambda \log(Q/P)}] = \log \mathbb{E}_P[e^{(\lambda+1)\log(Q/P)}] = \psi_P(\lambda+1) \\ \Rightarrow \quad &\psi_Q^*(\theta) = \sup_{\lambda\in\mathbb{R}} \theta\lambda - \psi_P(\lambda+1) = \psi_P^*(\theta) - \theta \end{aligned}$$

thus (E_0, E_1) in (13.1) is achievable.

Converse: We want to show that any achievable (E_0, E_1) pair must be below the curve $(E_0(\theta), E_1(\theta))$ in the above Neyman-Pearson test with parameter θ. Apply the strong converse bound we have:

$$\begin{aligned} &2^{-nE_0} + 2^{-n\theta}2^{-nE_1} \geq 2^{-n\psi_P^*(\theta)+o(n)} \\ \Rightarrow &\min(E_0, E_1 + \theta) \leq \psi_P^*(\theta), \ \forall n, \theta, -D(P\|Q) \leq \theta \leq D(Q\|P) \\ \Rightarrow &\text{ either } E_0 \leq \psi_P^*(\theta) \text{ or } E_1 \leq \psi_P^*(\theta) - \theta \end{aligned}$$

□

13.2 Equivalent forms of Theorem 13.1

Alternatively, the optimal (E_0, E_1)-tradeoff can be stated in the following equivalent forms:

Theorem 13.2. *1. The optimal exponents are given (parametrically) in terms of $\lambda \in [0,1]$ as*

$$E_0 = D(P_\lambda\|P), \qquad E_1 = D(P_\lambda\|Q) \tag{13.2}$$

where the distribution P_λ is tilting of P along T, cf. (12.14), *which moves from $P_0 = P$ to $P_1 = Q$ as λ ranges from 0 to 1:*

$$dP_\lambda = (dP)^{1-\lambda}(dQ)^\lambda \exp\{-\psi_P(\lambda)\}$$

2. Yet another characterization of the boundary is

$$E_1^*(E_0) = \min_{Q':D(Q'\|P)\le E_0} D(Q'\|Q)\,, \qquad 0 \le E_0 \le D(Q\|P) \tag{13.3}$$

Proof. The first part is verified trivially. Indeed, if we fix λ and let $\theta(\lambda) \triangleq \mathbb{E}_{P_\lambda}[T]$, then from (11.13) we have

$$D(P_\lambda\|P) = \psi_P^*(\theta)\,,$$

whereas

$$D(P_\lambda\|Q) = \mathbb{E}_{P_\lambda}[\log \frac{dP_\lambda}{dQ}] = \mathbb{E}_{P_\lambda}[\log \frac{dP_\lambda}{dP}\frac{dP}{dQ}] = D(P_\lambda\|P) - \mathbb{E}_{P_\lambda}[T] = \psi_P^*(\theta) - \theta\,.$$

Also from (11.12) we know that as λ ranges in $[0,1]$ the mean $\theta = \mathbb{E}_{P_\lambda}[T]$ ranges from $-D(P\|Q)$ to $D(Q\|P)$.

To prove the second claim (13.3), the key observation is the following: Since Q is itself a tilting of P along T (with $\lambda = 1$), the following two families of distributions

$$dP_\lambda = \exp\{\lambda T - \psi_P(\lambda)\} \cdot dP \tag{13.4}$$

$$dQ_{\lambda'} = \exp\{\lambda' T - \psi_Q(\lambda')\} \cdot dQ \tag{13.5}$$

are in fact the same family with $Q_{\lambda'} = P_{\lambda'+1}$.

Now, suppose that Q^* achieves the minimum in (13.3) and that $Q^* \ne Q$, $Q^* \ne P$ (these cases should be verified separately). Note that we have not shown that this minimum is achieved, but it will be clear that our argument can be extended to the case of when Q'_n is a sequence achieving the infimum. Then, on one hand, obviously

$$D(Q^*\|Q) = \min_{Q':D(Q'\|P)\le E_0} D(Q'\|Q) \le D(P\|Q)$$

On the other hand, since $E_0 \le D(Q\|P)$ we also have

$$D(Q^*\|P) \le D(Q\|P)\,.$$

Therefore,

$$\mathbb{E}_{Q^*}[T] = \mathbb{E}_{Q^*}[\log \frac{dQ^*}{dP}\frac{dQ}{dQ^*}] = D(Q^*\|P) - D(Q^*\|Q) \in [-D(P\|Q), D(Q\|P)]\,. \tag{13.6}$$

Next, we have from Corollary 12.1 that there exists a unique P_λ with the following three properties:[1]

$$\mathbb{E}_{P_\lambda}[T] = \mathbb{E}_{Q^*}[T] \tag{13.7}$$

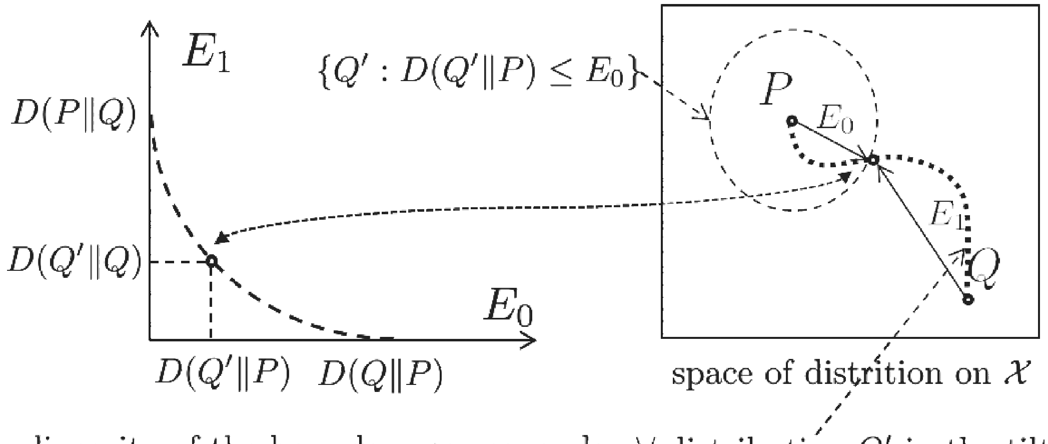
$$D(P_\lambda\|P) \le D(Q^*\|P) \tag{13.8}$$

$$D(P_\lambda\|Q) \le D(Q^*\|Q) \tag{13.9}$$

Thus, we immediately conclude that minimization in (13.3) can be restricted to Q^* belonging to the family of tilted distributions $\{P_\lambda, \lambda \in \mathbb{R}\}$. Furthermore, from (13.6) we also conclude that $\lambda \in [0,1]$. Hence, characterization of $E_1^*(E_0)$ given by (13.2) coincides with the one given by (13.3). □

[1] Small subtlety: In Corollary 12.1 we ask $\mathbb{E}_{Q^*}[T] \in (A,B)$. But A, B – the essential range of T – depend on the distribution under which the essential range is computed, cf. (12.10). Fortunately, we have $Q \ll P$ and $P \ll Q$, so essential range is the same under both P and Q. And furthermore (13.6) implies that $\mathbb{E}_{Q^*}[T] \in (A,B)$.

Note: Geometric interpretation of (13.3) is as follows: As λ increases from 0 to 1, or equivalently, θ increases from $-D(P\|Q)$ to $D(Q\|P)$, the optimal distribution traverses down the curve. This curve is in essense a geodesic connecting P to Q and exponents E_0,E_1 measure distances to P and Q. It may initially sound strange that the sum of distances to endpoints actually varies along the geodesic, but it is a natural phenomenon: just consider the unit circle with metric induced by the ambient Euclidean metric. Than if p and q are two antipodal points, the distance from intermediate point to endpoints do not sum up to $d(p,q) = 2$.



Non-linearity of the boundary corresponds to the scenario when the triangle inequality is not "="

$\forall$ distribution Q' in the tilted family, it minimizes E_0, E_1 simultaneously.
$\exists$ a unique optimal path from P to Q

13.3* Sequential Hypothesis Testing

Review: Filtrations, stopping times

- A sequence of nested σ-algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_n \cdots \subset \mathcal{F}$ is called a filtration of $\mathcal{F}$.
- A random variable τ is called a stopping time of a filtration $\mathcal{F}_n$ if a) τ is valued in $\mathbb{Z}_+$ and b) for every $n \ge 0$ the event $\{\tau \le n\} \in \mathcal{F}_n$.
- The σ-algebra $\mathcal{F}_\tau$ consists of all events E such that $E \cap \{\tau \le n\} \in \mathcal{F}_n$ for all $n \ge 0$.
- When $\mathcal{F}_n = \sigma\{X_1, \dots, X_n\}$ the interpretation is that τ is a time that can be determined by causally observing the sequence X_j, and random variables measurable with respect to $\mathcal{F}_\tau$ are precisely those whose value can be determined on the basis of knowing $(X_1, \dots, X_\tau)$.
- Let M_n be a martingale adapted to $\mathcal{F}_n$, i.e. M_n is $\mathcal{F}_n$-measurable and $\mathbb{E}[M_n|\mathcal{F}_k] = M_{\min(n,k)}$. Then $\tilde{M}_n = M_{\min(n,\tau)}$ is also a martingale. If collection $\{M_n\}$ is uniformly integrable then
$$\mathbb{E}[M_\tau] = \mathbb{E}[M_0]\,.$$
- For more details, see [Ç11, Chapter V].

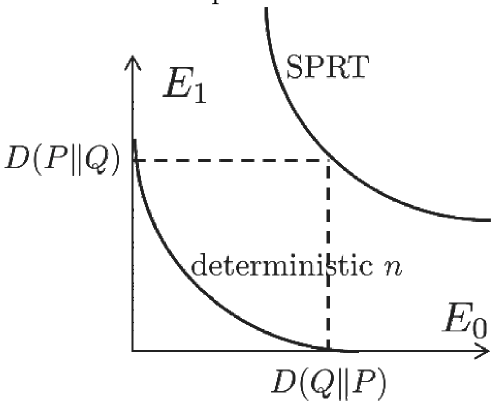
Different realizations of X_k are informative to different levels, the total "information" we receive follows a random process. Therefore, instead of fixing the sample size n, we can make n a stopping time τ, which gives a "better" (E_0, E_1) tradeoff. Solution is the concept of **sequential test**:

- Informally: Sequential test Z at each step declares either "H_0", "H_1" or "give me one more sample".
- Rigorous definition is as follows: A sequential hypothesis test is a stopping time τ of the filtration $\mathcal{F}_n \triangleq \sigma\{X_1, \dots, X_n\}$ and a random variable $Z \in \{0,1\}$ measurable with respect to $\mathcal{F}_\tau$.
- Each sequential test has the following performance metrics:

$$\alpha = \mathbb{P}[Z = 0], \qquad \beta = \mathbb{Q}[Z = 0] \tag{13.10}$$

$$l_0 = \mathbb{E}_{\mathbb{P}}[\tau], \qquad l_1 = \mathbb{E}_{\mathbb{Q}}[\tau] \tag{13.11}$$

The easiest way to see why sequential tests may be dramatically superior to fixed-sample size tests is the following example: Consider $P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ and $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_{-1}$. Since $P \not\perp Q$, we also have $P^n \not\perp Q^n$. Consequently, no finite-sample-size test can achieve zero error rates under both hypotheses. However, an obvious sequential test (wait for the first appearance of ±1) achieves zero error probability with finite average number of samples (2) under both hypotheses. This advantage is also seem very clearly in achievable error exponents.



Theorem 13.3. *Assume bounded LLR:*[2]

$$\left|\log \frac{P(x)}{Q(x)}\right| \le c_0, \forall x$$

where c_0 is some positive constant. If the error probabilities satisfy:

$$\pi_{1|0} \le 2^{-l_0 E_0}, \qquad \pi_{0|1} \le 2^{-l_1 E_1}$$

for large l_0, l_1, then the following inequality for the exponents holds

$$E_0 E_1 \le D(P\|Q) D(Q\|P).$$

[2] This assumption is satisfied for discrete distributions on finite spaces.

with optimal boundary achieved by the sequential probability ratio test SPRT(A,B) *(A, B are large positive numbers) defined as follows:*

$$\tau = \inf\{n : S_n \ge B \ \textit{or}\ S_n \le -A\}$$
$$Z = \begin{cases} 0, & \textit{if } S_\tau \ge B \\ 1, & \textit{if } S_\tau < -A \end{cases}$$

where

$$S_n = \sum_{k=1}^{n} \log \frac{P(X_k)}{Q(X_k)}$$

is the log likelihood function of the first k observations.

Note: (Intuition on SPRT) Under the usual hypothesis testing setup, we collect n samples, evaluate the LLR S_n, and compare it to the threshold to give the optimal test. Under the sequential setup with iid data, $\{S_n : n \ge 1\}$ is a *random walk*, which has positive (resp. negative) drift $D(P\|Q)$ (resp. $-D(Q\|P)$) under the null (resp. alternative)! SPRT test simply declare P if the random walk crosses the upper boundary B, or Q if the random walk crosses the upper boundary $-A$.

Proof. As preparation we show two useful identities:

- For any stopping time with $\mathbb{E}_P[\tau] < \infty$ we have

$$\mathbb{E}_P[S_\tau] = \mathbb{E}_P[\tau] D(P\|Q) \tag{13.12}$$

  and similarly, if $\mathbb{E}_Q[\tau] < \infty$ then

$$\mathbb{E}_Q[S_\tau] = -\mathbb{E}_Q[\tau] D(Q\|P)\,.$$

  To prove these, notice that

$$M_n = S_n - nD(P\|Q)$$

  is clearly a martingale w.r.t. $\mathcal{F}_n$. Consequently,

$$\tilde{M}_n \triangleq M_{\min(\tau,n)}$$

  is also a martingale. Thus

$$\mathbb{E}[\tilde{M}_n] = \mathbb{E}[\tilde{M}_0] = 0\,,$$

  or, equivalently,

$$\mathbb{E}[S_{\min(\tau,n)}] = \mathbb{E}[\min(\tau,n)] D(P\|Q)\,. \tag{13.13}$$

  This holds for every $n \ge 0$. From boundedness assumption we have $|S_n| \le nc$ and thus $|S_{\min(n,\tau)}| \le n\tau$, implying that collection $\{S_{\min(n,\tau)}, n \ge 0\}$ is uniformly integrable. Thus, we can take $n \to \infty$ in (13.13) and interchange expectation and limit safely to conclude (13.12).

- Let τ be a stopping time. The Radon-Nikodym derivative of $\mathbb{P}$ w.r.t. $\mathbb{Q}$ on σ-algebra $\mathcal{F}_\tau$ is given by

$$\frac{d\mathbb{P}|_{\mathcal{F}_\tau}}{d\mathbb{Q}|_{\mathcal{F}_\tau}} = \exp\{S_\tau\}\,.$$

  Indeed, what we need to verify is that for every event $E \in \mathcal{F}_\tau$ we have

$$\mathbb{E}_P[1_E] = \mathbb{E}_Q[\exp\{S_\tau\} 1_E] \tag{13.14}$$

To that end, consider a decomposition

$$1_E = \sum_{n\ge 0} 1_{E\cap\{\tau=n\}}\,.$$

By monotone convergence theorem applied to (13.14) it is sufficient to verify that for every n

$$\mathbb{E}_P[1_{E\cap\{\tau=n\}}] = \mathbb{E}_Q[\exp\{S_\tau\}1_{E\cap\{\tau=n\}}]\,. \tag{13.15}$$

This, however, follows from the fact that $E\cap\{\tau=n\}\in\mathcal{F}_n$ and $\frac{d\mathbb{P}|_{\mathcal{F}_n}}{d\mathbb{Q}|_{\mathcal{F}_n}} = \exp\{S_n\}$ by the very definition of S_n.

We now proceed to the proof. For **achievability** we apply (13.14) to infer

$$\begin{aligned}\pi_{1|0} &= \mathbb{P}[S_\tau \le -A]\\ &= \mathbb{E}_Q[\exp\{S_\tau\}1\{S_\tau\le -A\}]\\ &\le e^{-A}\end{aligned}$$

Next, we denot $\tau_0 = \inf\{n : S_n \ge B\}$ and observe that $\tau\le\tau_0$, whereas expectation of τ_0 we estimate from (13.12):

$$\mathbb{E}_P[\tau] \le \mathbb{E}_P[\tau_0] = \mathbb{E}_P[S_{\tau_0}] \le B + c_0\,,$$

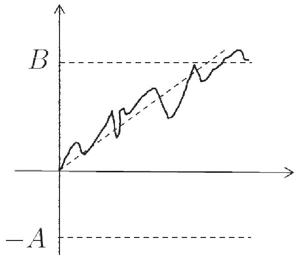
where in the last step we used the boundedness assumption to infer

$$S_{\tau_0} \le B + c_0$$

Thus

$$l_0 = \mathbb{E}_{\mathbb{P}}[\tau] \le \mathbb{E}_{\mathbb{P}}[\tau_0] \le \frac{B+c_0}{D(P\|Q)} \approx \frac{B}{D(P\|Q)} \text{ for large } B$$

Similarly we can show $\pi_{0|1}\le e^{-B}$ and $l_1 \le \frac{A}{D(Q\|P)}$ for large A. Take $B = l_0 D(P\|Q)$, $A = l_1 D(Q\|P)$, this shows the achievability.



under P, $S_n - nD(P\|Q)$ is a martingale

Converse: Assume (E_0, E_1) achievable for large l_0, l_1 and apply data processing inequality of divergence:

$$\begin{aligned} d(\mathbb{P}(Z=1)\|\mathbb{Q}(Z=1)) &\le D(\mathbb{P}\|\mathbb{Q})\big|_{\mathcal{F}_\tau}\\ &= \mathbb{E}_P[S_\tau] \qquad\qquad = \mathbb{E}_{\mathbb{P}}[\tau]D(P\|Q) \quad \text{from (13.12)}\\ &= l_0 D(P\|Q)\end{aligned}$$

notice that for l_0E_0 and l_1E_1 large, we have $d(\mathbb{P}(Z=1)\|\mathbb{Q}(Z=1)) \approx l_1E_1$, therefore $l_1E_1 \lesssim l_0 D(P\|Q)$. Similarly we can show that $l_0E_0 \lesssim l_1 D(Q\|P)$, finally we have

$$E_0E_1 \le D(P\|Q)D(Q\|P), \text{ as } l_0, l_1 \to \infty$$

□