Minimax Lower Bounds

Benedikt Koch

July 2026

About These Notes

These notes are based primarily on my reading of Chapter 2 of Tsybakov (2009). They slightly generalize some of the results, present some proofs from a more information-theoretic perspective, include additional examples, and (hopefully) are more intuitive to the reader. The presentation is self-contained but has not been peer reviewed. Any errors are my own.

Contents

1 Setup

To formulate a minimax estimation problem, we specify the following ingredients:

- A *parameter space* Θ containing the unknown true parameter $\theta^\star \in \Theta$. For example, Θ may be a class of β-Hölder functions or a subset of $\mathbb{R}^q$ representing the possible mean vectors of a q-dimensional Gaussian distribution.
- An *estimand* $g(\theta^\star)$ for a measurable $g : \Theta \to \mathcal{G}$ where $\mathcal{G}$ is the target space.
- A *statistical model* $\mathcal{P}^{(n)} = \{P_\theta^{(n)} : \theta \in \Theta\}$, where each $P_\theta^{(n)}$ is a candidate joint distribution for n observations on the common measurable sample space $(\mathcal{X}^n, \mathcal{A}^{\otimes n})$.
- An observable *dataset* $X^{(n)} = (X_1, \ldots, X_n)$, where $X_i \in \mathcal{X}$. Under the true parameter, $X^{(n)} \sim P_{\theta^\star}^{(n)}$. The observations need not be independent or identically distributed.
- A *semi-distance* $d : \mathcal{G} \times \mathcal{G} \to [0, \infty)$ for measuring the estimation error.
- A *loss function* $\ell : [0, \infty) \to [0, \infty)$ that is non-decreasing, satisfies $\ell(0) = 0$, and is not identically zero. For example $\ell(u) = u^p, p > 0$ or $\ell(u) = \mathbf{1}\{u \geq A\}, A > 0$.

An estimator of $g(\theta^\star)$ is a measurable function $T_n = T_n(X^{(n)}) \in \mathcal{G}$. Because $\theta^\star$ is unknown and is only known to belong to Θ, we assess T_n through its *maximal risk*

$$r_{\ell,n}(T_n; \Theta) := \sup_{\theta \in \Theta} \mathbb{E}_\theta^{(n)}[\ell(d(T_n, g(\theta)))].$$

Our goal is to quantify how good an estimator can be when assessed by the maximum risk. Since the performance depends on the size of the dataset n, we normalize it by a positive sequence $\psi_n \to 0$ called a candidate *rate of convergence*. The minimax-lower-bound problem is to find ψ_n to establish

$$\liminf_{n \to \infty} \inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta^{(n)}[\ell(\psi_n^{-1} d(T_n, g(\theta)))] \geq c, \tag{1}$$

for some constant $c > 0$ independent of n, where the infimum is taken over all measurable estimators based on n observations. This inequality shows that no estimator can converge uniformly over Θ at a rate faster than ψ_n.

2 From Estimation to Testing

The minimax estimation problem in Equation (1) can be reduced to a testing problem. By Markov inequality for any $A > 0$ with $\ell(A) > 0$

$$\begin{aligned}
\inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta^{(n)}[\ell(\psi_n^{-1} d(T_n, g(\theta)))] &\geq \ell(A) \inf_{T_n} \sup_{\theta \in \Theta} P_\theta^{(n)}(\psi_n^{-1} d(T_n, g(\theta)) \geq A) \\
&= \ell(A) \inf_{T_n} \sup_{\theta \in \Theta} P_\theta^{(n)}(d(T_n, g(\theta)) \geq s_n) \\
&\geq \ell(A) \inf_{T_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta^{(n)}(d(T_n, g(\theta)) \geq s_n). \\
&= \ell(A) \inf_{T_n} \max_{0 \leq j \leq M} P_{\theta_j}^{(n)}(d(T_n, g(\theta_j)) \geq s_n).
\end{aligned}$$

where $s_n = A\psi_n$ and where we fixed a finite collection of hypotheses $\{\theta_0, \ldots, \theta_M\} \subseteq \Theta$. The hypotheses $\theta_j = \theta_{j,n}$, as well as $M = M_n$, may depend on n. Our aim is to choose the hypotheses so that their target values are $2s_n$-separated, i.e.,

$$\Delta_{n,M} := \min_{0 \leq j < k \leq M} d(g(\theta_j), g(\theta_k)) \geq 2s_n = 2A\psi_n, \tag{2}$$

for all sufficiently large n. A convenient sufficient condition is $\liminf_{n \to \infty} \frac{\Delta_{n,M}}{\psi_n} = c_\Delta > 0$. In this case, choose A such that $\ell(A) > 0$ and $A \geq \frac{c_\Delta}{2}$ (if exists).

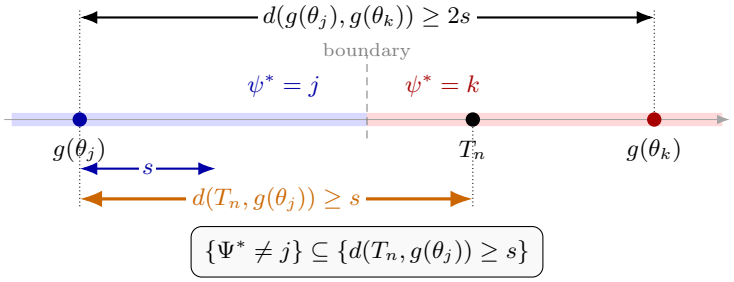

Figure 1: One-dimensional illustration of the estimation-to-testing reduction. If the true target is $g(\theta_j)$ but the minimum-distance test selects $k \neq j$, then the estimator must be at least distance s from $g(\theta_j)$.

A *test* is a measurable function $\Psi_n : \mathcal{X}^{(n)} \to \{0, 1, \ldots, M\}$. If Equation (2) holds, then for any estimator T_n

$$P_{\theta_j}(d(T_n, g(\theta_j)) \geq s_n) \geq P_{\theta_j}(\Psi^*_{T_n} \neq j), \quad j = 0, 1, \ldots, M \tag{3}$$

where $\Psi^*_{T_n} = \arg\min_{0 \leq k \leq M} d(T_n, g(\theta_k))$ is the *minimum distance test* based on T_n. See Figure 1 for a justification of the inequality in Equation (3).

Putting the pieces together, as long as Equation (2) holds, we have for $s_n = A\psi_n$

$$\begin{aligned}
\inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}^{(n)}_\theta[\ell(\psi_n^{-1} d(T_n, g(\theta)))] &\geq \ell(A) \inf_{T_n} \sup_{\theta \in \Theta} P^{(n)}_\theta(d(T_n, g(\theta)) \geq s_n) \\
&\geq \ell(A) \inf_{T_n} \max_{0 \leq j \leq M} P^{(n)}_{\theta_j}(d(T_n, g(\theta_j)) \geq s_n) \\
&\geq \ell(A) \inf_{T_n} \max_{0 \leq j \leq M} P^{(n)}_{\theta_j}(\Psi^*_{T_n} \neq j) \\
&\geq \ell(A) \inf_{\Psi_n} \max_{0 \leq j \leq M} P^{(n)}_{\theta_j}(\Psi_n \neq j),
\end{aligned} \tag{4}$$

where the infimum is taken over all tests (based on n samples). Thus, we have transformed the estimation problem into a testing problem. The lower bound is large, if the testing problem is difficult. Intuitively, this means that we want $\{P_{\theta_j}\}_{j=1}^M$ to be *close* to each other, which is the subject of the next section.

3 Lower Bounds

3.1 Lower Bounds based on Two Hypothesis

Often the simple case $M = 1$ is enough. This means that we take only two hypotheses θ_0 and θ_1 belonging to Θ and write $P^{(n)}_j := P^{(n)}_{\theta_j}$ for $j \in \{0, 1\}$. this case, Equation (4) involves

$$\begin{aligned}
\inf_{\Psi_n} \max_{j=0,1} P^{(n)}_j(\Psi_n \neq j) &= \inf_{\Psi_n} \max\{P^{(n)}_0(\Psi_n = 1), P^{(n)}_1(\Psi_n = 0)\} \\
&\geq \frac{1}{2} \inf_{\Psi_n}(P^{(n)}_0(\Psi_n = 1) + P^{(n)}_1(\Psi_n = 0)) \\
&= \frac{1}{2} \inf_{\Psi_n}(1 - P^{(n)}_0(\Psi_n = 0) + P^{(n)}_1(\Psi_n = 0)) \\
&= \frac{1}{2} \inf_{\Psi_n}(1 - [P^{(n)}_0(\Psi_n = 0) - P^{(n)}_1(\Psi_n = 0)]) \\
&\geq \frac{1}{2} \inf_{\Psi_n}(1 - |P^{(n)}_0(\Psi_n = 0) - P^{(n)}_1(\Psi_n = 0)|) \\
&\geq \frac{1}{2}(1 - V(P^{(n)}_0, P^{(n)}_1)),
\end{aligned} \tag{5}$$

where $V(\cdot,\cdot)$ is the total variation distance (see Appendix A.1). Note that the bounds *after* Equation (5) are *sharp*. Indeed, the infimum of the *sum* of the two error probabilities is attained by the maximum-likelihood test

$$\Psi_{\mathrm{ML},n} := \begin{cases} 0, & p_0^{(n)} \geq p_1^{(n)}, \\ 1, & p_0^{(n)} < p_1^{(n)}. \end{cases}$$

Here $p_j^{(n)} = dP_j^{(n)}/d\nu$ for any measure ν dominating both joint laws, for example $\nu = P_0^{(n)} + P_1^{(n)}$. We further have

$$P_0^{(n)}(\Psi_{\mathrm{ML},n} = 1) + P_1^{(n)}(\Psi_{\mathrm{ML},n} = 0) = 1 - V(P_0^{(n)}, P_1^{(n)}).$$

We can summarize our results so far in the following theorem.

Theorem 1 (Le Cam's two-point lower bound, Theorem 2.2, Tsybakov (2009)) *For each n, let $\theta_{0,n}, \theta_{1,n} \in \Theta$. Suppose that there exist $A_n > 0$ and $\alpha_n \in [0,1)$ such that*

1. $d\big(g(\theta_{0,n}), g(\theta_{1,n})\big) \geq 2A_n\psi_n$ *and* $\ell(A_n) > 0$;
2. $V\big(P_{\theta_{0,n}}^{(n)}, P_{\theta_{1,n}}^{(n)}\big) \leq \alpha_n$.

Then

$$\inf_{T_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta^{(n)} \left[\ell\left(\psi_n^{-1} d(T_n, g(\theta))\right)\right] \geq \frac{\ell(A_n)}{2}(1-\alpha_n).$$

Theorem 1 captures the balance underlying the method. The hypotheses must be sufficiently separated,

$$d\big(g(\theta_{0,n}), g(\theta_{1,n})\big) \geq 2A_n\psi_n,$$

while the corresponding joint laws must remain sufficiently close,

$$V\big(P_{\theta_{0,n}}^{(n)}, P_{\theta_{1,n}}^{(n)}\big) \leq \alpha_n.$$

One may set

$$A_n := \frac{d\big(g(\theta_{0,n}), g(\theta_{1,n})\big)}{2\psi_n},$$

as long as $\ell(A_n)$ is bounded from below for sufficiently large n, so that the separation condition holds with equality (strongest possible separation). It then suffices to choose ψ_n and the two hypotheses such that, for all sufficiently large n,

$$\frac{\ell(A_n)}{2}(1-\alpha_n) \geq c > 0.$$

Theorem 1 expresses the lower bound in terms of total variation distance. The following lemma allows us to verify the required closeness condition using other standard probability distances.

Lemma 1 (Comparison of probability distances) *Let P and Q be probability measures on the same measurable space. Denote by $V(P,Q)$ the* total variation distance *(Appendix A.1), by $H(P,Q)$ the* Hellinger distance *(Appendix A.2), by $K(P,Q)$ the* Kullback–Leibler divergence *(Appendix A.3), and by $\chi^2(P,Q)$ the* chi-squared divergence *(Appendix A.4). Then (Tsybakov, 2009, Equation (2.27))*

$$V(P,Q) \leq H(P,Q) \leq \sqrt{K(P,Q)} \leq \sqrt{\chi^2(P,Q)}. \tag{6}$$

Moreover, Pinsker's inequality gives the sharper direct comparison (Tsybakov, 2009, Lemma 2.5)

$$V(P,Q) \leq \sqrt{\frac{1}{2}K(P,Q)}. \tag{7}$$

Consequently, the condition $V(P,Q) \leq \alpha$ in Theorem 1 is guaranteed, for example, by any of the conditions

$$H(P,Q) \leq \alpha, \qquad K(P,Q) \leq 2\alpha^2, \qquad \text{or} \qquad \chi^2(P,Q) \leq \alpha^2.$$

While lower bounds based on two-hypotheses are convenient, they are not always sufficient, as the following example shows.

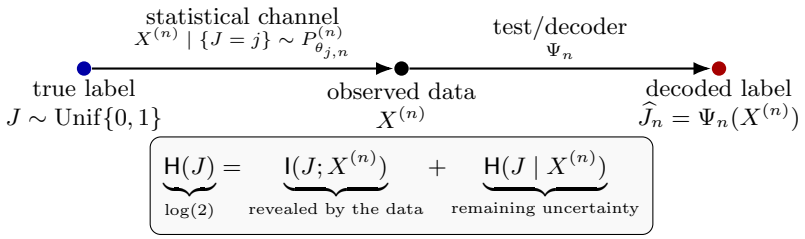

Figure 2: Information-theoretic view of binary hypothesis testing. Nature selects J, the statistical channel generates $X^{(n)}$, and the test Ψ_n returns the decoded label $\widehat{J}_n$.

Example 1 (Gaussian Means with Growing Dimension) *Let $q = q_n \to \infty$ with $q_n = o(n)$, and consider*

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\theta, I_q), \qquad \theta \in \Theta_n = \mathbb{R}^q.$$

We estimate $g(\theta) = \theta$ using $d(\theta, \theta') = \|\theta - \theta'\|_2$ and $\ell(u) = u^2$. The sample mean attains the minimax risk

$$\sup_{\theta \in \Theta_n} \mathbb{E}^{(n)}_\theta \left[\left\| \bar{X}_n - \theta \right\|_2^2 \right] = \frac{q}{n},$$

which turns out to be optimal. Thus, set $\psi_n = \sqrt{\frac{q}{n}}$. Now choose any two hypotheses $\theta_{0,n}$ and $\theta_{1,n}$. Then

$$\alpha_n := V\left(P^{(n)}_{\theta_{0,n}}, P^{(n)}_{\theta_{1,n}}\right) = 1 - 2\Phi(-t_n),$$

where $t_n := \frac{\sqrt{n}\Delta_n}{2}$ and $\Delta_n := \|\theta_{1,n} - \theta_{0,n}\|_2$. The strongest admissible separation constant in Theorem 1 is

$$A_n = \frac{\Delta_n}{2\psi_n} = \frac{t_n}{\sqrt{q}}.$$

Hence, the lower bound in Theorem 1 becomes

$$\frac{\ell(A_n)}{2}(1 - \alpha_n) = \frac{t_n^2 \Phi(-t_n)}{q} \le \frac{C_0}{q}, \qquad C_0 := \sup_{t \ge 0} t^2 \Phi(-t) < \infty,$$

which vanishes as $q = q(n) \to \infty$ as $n \to \infty$. Therefore, every two-hypothesis lower bound converges to zero after normalization by $\psi_n^2 = q/n$. Equivalently, on the original squared-risk scale, it yields only a lower bound of order $1/n$, whereas the true minimax risk is of order q/n.

3.2 Lower Bounds based on Many Hypothesis

We discuss Fano's lemma (general finite hypothesis sets controlled through Kullback–Leibler divergence), Assouad's lemma (hypercube-indexed hypotheses controlled through local probability-distance comparisons), and the method of fuzzy hypotheses.

3.2.1 Fano's Lemma

The limitation of the two-hypothesis method in Example 1 has an information-theoretic interpretation. Let J be uniformly distributed on $\{0,1\}$, and let $X^{(n)} \mid \{J = j\} \sim P^{(n)}_{\theta_{j,n}}$. Here, J indexes the true hypothesis selected by nature, which is encoded through the statistical channel $J \to X^{(n)}$ that sends the data $X^{(n)}$. After observing the data, a test Ψ_n attempts to decode the label through $\widehat{J}_n = \Psi_n(X^{(n)})$, as illustrated in Figure 2.

Before observing the data, uncertainty about the true hypothesis index J is quantified by its *entropy*, $\mathsf{H}(J) = \log(2)$ (see Appendix B.1). Observing $X^{(n)}$ reveals an amount of information quantified by the *mutual information* $\mathsf{I}(J; X^{(n)})$ (see Appendix B.3). The remaining uncertainty is the *conditional entropy* (see Appendix B.2),

$$\mathsf{H}(J \mid X^{(n)}) = \mathsf{H}(J) - \mathsf{I}(J; X^{(n)}) = \log(2) - \mathsf{I}(J; X^{(n)}).$$

Intuitively, in order to derive a meaningful minimax lower bound, the conditional entropy must be bounded away from zero. In this case, even the best decoder (test) cannot recover the true label.

In Example 1, by Lemma 2,

$$\begin{aligned}\mathsf{H}(J \mid X^{(n)}) &\geq \log(2) - \frac{1}{2} K\left(P_{\theta_{1,n}}^{(n)}, P_{\theta_{0,n}}^{(n)}\right) \\ &= \log(2) - \frac{n}{4}\left\|\theta_{1,n} - \theta_{0,n}\right\|_2^2. \end{aligned} \tag{8}$$

Thus, to ensure that $\mathsf{H}(J \mid X^{(n)})$ remains bounded away from zero, we must have $\|\theta_{1,n} - \theta_{0,n}\|_2^2 \lesssim 1/n$.

However, since we know that $\psi_n = \sqrt{q/n}$, the separation condition in Theorem 1 requires

$$\|\theta_{1,n} - \theta_{0,n}\|_2^2 \geq 4A_n^2 \frac{q}{n}.$$

Since under the square loss we must have $A_n \geq c > 0$ for sufficiently large n, the squared separation must satisfy $\|\theta_{1,n} - \theta_{0,n}\|_2^2 \gtrsim q/n$. However, this yields vanishing uncertainty in Equation (8), which allows the true label to be recovered.

The limitation above is that two hypotheses provide only $\mathsf{H}(J) = \log(2)$ units of initial uncertainty. The natural remedy is to use $M+1$ hypotheses, so that $\mathsf{H}(J) = \log(M+1)$ can grow with the complexity of the problem. If this initial uncertainty dominates the information revealed by the data, no test can reliably decode J. Fano's lemma makes this intuition precise.

In the following, for hypotheses $\theta_{0,n}, \ldots, \theta_{M,n}$, write $P_{j,n} := P_{\theta_{j,n}}^{(n)}$. As in the previous section, we proceed with Equation (4) by bounding the maximum by the average

$$\begin{aligned}\inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta^{(n)}[\ell(\psi_n^{-1} d(T_n, g(\theta)))] &\geq \ell(A) \inf_{\Psi_n} \max_{0 \leq j \leq M} P_{j,n}(\Psi_n \neq j) \\ &\geq \ell(A) \inf_{\Psi_n} \frac{1}{M+1} \sum_{j=0}^{M} P_{j,n}(\Psi_n \neq j). \end{aligned} \tag{9}$$

Theorem 2 (Fano's many-hypothesis lower bound, Corollary 2.6, Tsybakov (2009)) *For each n, let $M = M_n \geq 2$ and let $\theta_{0,n}, \ldots, \theta_{M,n} \in \Theta$. Suppose that there exist $A_n > 0$ and $\alpha_n \in (0,1)$ such that*

1. $\min_{0 \leq j < k \leq M} d\big(g(\theta_{j,n}), g(\theta_{k,n})\big) \geq 2A_n\psi_n$ *and* $\ell(A_n) > 0$;
2. $\frac{1}{M+1}\sum_{j=1}^{M} K(P_{j,n}, P_{0,n}) \leq \alpha_n \log(M)$.

Then

$$\inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta^{(n)}\left[\ell\left(\psi_n^{-1} d(T_n, g(\theta))\right)\right] \geq \ell(A_n)\left[\frac{\log(M+1) - \log(2)}{\log(M)} - \alpha_n\right].$$

Proof: Let J be uniformly distributed on $\{0, 1, \ldots, M\}$ and, conditional on $J = j$, let

$$X^{(n)} \mid \{J = j\} \sim P_{j,n}.$$

Thus, J indexes the true hypothesis, $X^{(n)}$ denotes the observed data, and $\{P_{j,n}\}_{j=0}^{M}$ are the candidate data-generating distributions, each assigned equal prior probability. We quantify what the data reveal about the true hypothesis through the remaining uncertainty $\mathsf{H}(J \mid X^{(n)})$.

Fix an arbitrary test $\Psi_n : \mathcal{X}^{(n)} \to \{0, 1, \ldots, M\}$ and define its error indicator by $E_n := \mathbf{1}\big\{\Psi_n(X^{(n)}) \neq J\big\}$.

Its average error probability is

$$\begin{aligned} p &:= P(E_n = 1) \\ &= \sum_{j=0}^{M} P(E_n = 1 \mid J = j) P(J = j) \\ &= \frac{1}{M+1} \sum_{j=0}^{M} P(\Psi_n(X^{(n)}) \neq j \mid J = j) \\ &= \frac{1}{M+1} \sum_{j=0}^{M} P_{j,n}(\Psi_n(X^{(n)}) \neq j), \end{aligned}$$

which is the average testing error appearing in Equation (9).

Because E_n is determined by $(J, X^{(n)})$, $\mathsf{H}(E_n \mid J, X^{(n)}) = 0$. Applying the chain rule for conditional entropy (see Appendix B.2) twice therefore gives

$$\begin{aligned} \mathsf{H}(J \mid X^{(n)}) &= \mathsf{H}(E_n, J \mid X^{(n)}) - \mathsf{H}(E_n \mid J, X^{(n)}) \\ &= \mathsf{H}(E_n, J \mid X^{(n)}) \\ &= \mathsf{H}(E_n \mid X^{(n)}) + \mathsf{H}(J \mid E_n, X^{(n)}). \end{aligned} \tag{10}$$

Here, $\mathsf{H}(E_n \mid X^{(n)})$ measures the remaining uncertainty about whether the test is correct after observing the data. Since conditioning reduces entropy and E_n takes only two values,

$$\mathsf{H}(E_n \mid X^{(n)}) \leq \mathsf{H}(E_n) \leq \log(2). \tag{11}$$

Next, $\mathsf{H}(J \mid E_n, X^{(n)})$ measures the remaining uncertainty about the true hypothesis after observing the data and whether the test is correct. On the event $E_n = 0$, we have $J = \Psi_n(X^{(n)})$, so since $X^{(n)}$ is observed, J is known and thus

$$\mathsf{H}(J \mid E_n = 0, X^{(n)}) = 0.$$

On the event $E_n = 1$, the test output $\Psi_n(X^{(n)})$ is incorrect and therefore rules out one of the $M + 1$ possible values of J. Hence, J can take at most M remaining values, and therefore

$$\mathsf{H}(J \mid E_n = 1, X^{(n)}) \leq \log(M).$$

It follows that

$$\begin{aligned} \mathsf{H}(J \mid E_n, X^{(n)}) &= p\mathsf{H}(J \mid E_n = 1, X^{(n)}) + (1-p)\mathsf{H}(J \mid E_n = 0, X^{(n)}) \\ &= p\mathsf{H}(J \mid E_n = 1, X^{(n)}) \\ &= p\mathsf{H}(J \mid E_n = 1, X^{(n)}) \\ &\leq p \log(M). \end{aligned} \tag{12}$$

Combining Equations (10), (11), and (12) yields

$$\mathsf{H}(J \mid X^{(n)}) \leq \log(2) + p \log(M).$$

By the by the definition of mutual information (see Appendix B.3) and since J is uniformly distributed, $\mathsf{H}(J \mid X^{(n)}) = \mathsf{H}(J) - \mathsf{I}(J; X^{(n)}) = \log(M+1) - \mathsf{I}(J; X^{(n)})$. Consequently,

$$\log(M+1) - \mathsf{I}(J; X^{(n)}) \leq \log(2) + p \log(M).$$

Since $M \geq 2$, rearranging gives

$$p \geq \frac{\log(M+1) - \log(2) - \mathsf{I}(J; X^{(n)})}{\log(M)}.$$

By Lemma 2 and the assumed KL bound,

$$\mathsf{I}\left(J; X^{(n)}\right) \le \frac{1}{M+1}\sum_{j=1}^{M} K\left(P_{\theta_j}^{(n)}, P_{\theta_0}^{(n)}\right) \le \alpha_n \log(M).$$

Substituting this bound into the preceding inequality gives

$$p \ge \frac{\log(M+1) - \log(2)}{\log(M)} - \alpha_n.$$

Because Ψ_n was arbitrary, taking the infimum over all tests and using Equation (4) yields

$$\begin{aligned}
\inf_{T_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta^{(n)}[\ell(\psi_n^{-1} d(T_n, g(\theta)))] &\ge \ell(A_n) \inf_{\Psi_n} \max_{0\le j\le M} P_{j,n}(\Psi_n \ne j) \\
&\ge \ell(A_n) \inf_{\Psi_n} \frac{1}{M+1}\sum_{j=0}^{M} P_{j,n}(\Psi_n \ne j) \\
&\ge \ell(A_n)\left[\frac{\log(M+1)-\log(2)}{\log(M)} - \alpha_n\right]
\end{aligned}$$

□

If $M_n \to \infty$, $\limsup_{n\to\infty} \alpha_n \le \alpha < 1$, and $\liminf_{n\to\infty} \ell(A_n) > 0$, then

$$\liminf_{n\to\infty}\left[\frac{\log(M_n+1)-\log(2)}{\log(M_n)} - \alpha_n\right] \ge 1 - \alpha > 0.$$

Hence, the lower bound in Theorem 2 remains bounded away from zero. Given hypotheses $\{\theta_{0,n}, \ldots, \theta_{M_n,n}\}$, the appropriate choice is

$$A_n := \frac{1}{2\psi_n} \min_{0\le j<k\le M_n} d\big(g(\theta_{j,n}), g(\theta_{k,n})\big).$$

Lemma 2 (Average KL decomposition) *Let J be uniformly distributed on $\{0, \ldots, M\}$, and suppose that*

$$X^{(n)} \mid \{J = j\} \sim P_{\theta_j}^{(n)},$$

such that the marginal of $X^{(n)}$ is given by $\overline{P}_n = \frac{1}{M+1}\sum_{j=0}^{M} P_{\theta_j}^{(n)}$. Then, for any probability measure $Q^{(n)}$ for which the relevant KL divergences are finite,

$$\frac{1}{M+1}\sum_{j=0}^{M} K\left(P_{\theta_j}^{(n)}, Q^{(n)}\right) = \mathsf{I}\left(J; X^{(n)}\right) + K\left(\overline{P}_n, Q^{(n)}\right),$$

where I denotes the mutual information *(see Appendix B.3). Consequently,*

$$\mathsf{I}\left(J; X^{(n)}\right) \le \frac{1}{M+1}\sum_{j=0}^{M} K\left(P_{\theta_j}^{(n)}, Q^{(n)}\right).$$

In particular, taking $Q^{(n)} = P_{\theta_0}^{(n)}$ gives

$$\mathsf{I}\left(J; X^{(n)}\right) \le \frac{1}{M+1}\sum_{j=1}^{M} K\left(P_{\theta_j}^{(n)}, P_{\theta_0}^{(n)}\right).$$

Proof: Let p_j, $\overline{p}$, and q denote the densities of $P_{\theta_j}^{(n)}$, $\overline{P}_n$, and $Q^{(n)}$, respectively, with respect to a common dominating measure ν. Then

$$\overline{p} = \frac{1}{M+1}\sum_{j=0}^{M} p_j.$$

If

$$\mathsf{I}\left(J; X^{(n)}\right) = \frac{1}{M+1}\sum_{j=0}^{M}\int p_j \log\left(\frac{p_j}{\overline{p}}\right)\, d\nu, \tag{13}$$

then

$$\begin{aligned}
\frac{1}{M+1}\sum_{j=0}^{M} K\left(P_{\theta_j}^{(n)}, Q^{(n)}\right) &= \frac{1}{M+1}\sum_{j=0}^{M}\int p_j \log\left(\frac{p_j}{q}\right)\, d\nu \\
&= \frac{1}{M+1}\sum_{j=0}^{M}\int p_j \log\left(\frac{p_j}{\overline{p}}\right)\, d\nu + \frac{1}{M+1}\sum_{j=0}^{M}\int p_j \log\left(\frac{\overline{p}}{q}\right)\, d\nu \\
&= \mathsf{I}\left(J; X^{(n)}\right) + \int \overline{p}\log\left(\frac{\overline{p}}{q}\right)\, d\nu \\
&= \mathsf{I}\left(J; X^{(n)}\right) + K\left(\overline{P}_n, Q^{(n)}\right).
\end{aligned}$$

The inequalities follow from the nonnegativity of KL divergence. Thus it remains to show Equation (13).

By definition of mutual information (see Appendix B.3)

$$\mathsf{I}(J; X^{(n)}) = K\left(P_{J,X^{(n)}}, P_J \otimes P_{X^{(n)}}\right).$$

The joint density of $(J, X^{(n)})$ is

$$p_{J,X^{(n)}}(j, x) = p(x \mid J = j)P(J = j) = \frac{1}{M+1}p_j(x).$$

The density under the product is

$$p_J \otimes P_{X^{(n)}}(j, x) = p_J(j)p_{X^{(n)}}(x) = \frac{1}{M+1}\bar{p}(x).$$

Substituting these densities into the definition of mutual information gives

$$\begin{aligned}
\mathsf{I}(J; X^{(n)}) &= K\left(P_{J,X^{(n)}}, P_J \otimes P_{X^{(n)}}\right) \\
&= \sum_{j=0}^{M}\int_{\mathcal{X}^{(n)}} \frac{1}{M+1}p_j \log\left(\frac{p_j/(M+1)}{\bar{p}/(M+1)}\right)\, d\nu \\
&= \frac{1}{M+1}\sum_{j=0}^{J}\int_{\mathcal{X}^{(n)}} p_j \log\left(\frac{p_j}{\bar{p}}\right)\, d\nu.
\end{aligned}$$

□

Lemma 3 (Varshamov–Gilbert bound, Lemma 2.9, Tsybakov (2009)) *Let $q \geq 8$. For $\omega, \omega' \in \{0,1\}^q$, define the Hamming distance by*

$$\rho_H(\omega, \omega') := \sum_{r=1}^{q} \mathbf{1}\{\omega_r \neq \omega'_r\}.$$

Then there exist $\omega_0, \omega_1, \ldots, \omega_M \in \{0,1\}^q$, with $\omega_0 = 0$, such that

$$\rho_H(\omega_j, \omega_k) \geq \frac{q}{8}, \qquad 0 \leq j < k \leq M,$$

and

$$M \geq 2^{q/8}.$$

Consequently, $\log(M) \geq q\log(2)/8$.

Example 2 *We now use Theorem 2 to recover the sharp minimax rate. For $q \geq 8$, Lemma 3 yields $\omega_0, \dots, \omega_M \in \{0,1\}^q$, with $\omega_0 = 0$, such that*

$$M \geq 2^{q/8} \qquad \text{and} \qquad \rho_H(\omega_j, \omega_k) \geq \frac{q}{8}, \quad j \neq k,$$

where $\rho_H(\omega, \omega') := \sum_{r=1}^{q} \mathbf{1}\{\omega_r \neq \omega'_r\}$ is the Hamming distance. Set

$$a^2 := \frac{\log(2)}{8}, \qquad \theta_{j,n} := \frac{a}{\sqrt{n}}\omega_j, \qquad j = 0, \dots, M.$$

Then, for $j \neq k$,

$$\|\theta_{j,n} - \theta_{k,n}\|_2^2 = \frac{a^2}{n}\rho_H(\omega_j, \omega_k) \geq \frac{a^2 q}{8n} = 4A^2\psi_n^2, \qquad A := \frac{a}{4\sqrt{2}}.$$

Moreover, since $\theta_{0,n} = 0$ and $\log(M) \geq q\log(2)/8$,

$$\frac{1}{M+1}\sum_{j=1}^{M} K(P_{j,n}, P_{0,n}) \leq \frac{a^2 q}{2} = \frac{q\log(2)}{16} \leq \frac{1}{2}\log(M).$$

Thus, Theorem 2, with $A_n = A$ and $\alpha_n = 1/2$, gives

$$\frac{n}{q}\inf_{T_n}\sup_{\theta\in\Theta_n} \mathbb{E}_\theta^{(n)}\left[\|T_n - \theta\|_2^2\right] \geq \frac{\log(2)}{256}\left[\frac{\log(M+1) - \log(2)}{\log(M)} - \frac{1}{2}\right].$$

Because $q = q_n \to \infty$, we have $M = M_n \to \infty$, and hence

$$\liminf_{n\to\infty} \frac{n}{q_n}\inf_{T_n}\sup_{\theta\in\Theta_n} \mathbb{E}_\theta^{(n)}\left[\|T_n - \theta\|_2^2\right] \geq \frac{\log(2)}{512}.$$

Together with the upper bound attained by the sample mean in Example 1, this proves that the minimax risk is of order q_n/n.

3.2.2 Assouad's Lemma

The version of Fano's lemma above allows a general finite collection of hypotheses, but its application requires a globally separated packing and control of an average Kullback–Leibler divergence relative to a reference law. Assouad's lemma makes the complementary tradeoff. It restricts hypotheses to the vertices of a hypercube and reduces the problem to binary tests along neighboring edges. It admits a total variation version (which always exists) and is generally easier to verify.

Let

$$\Omega_m := \{0,1\}^m.$$

For $\omega \in \Omega_m$ and $k \in \{1, \dots, m\}$, let For $\omega = (\omega_1, \dots, \omega_m) \in \Omega_m$ and $k \in \{1, \dots, m\}$, let $\omega^{(k)} := (\omega_1, \dots, \omega_{k-1}, 1-\omega_k, \omega_{k+1}, \dots, \omega_m)$ denote the vertex obtained by flipping the kth coordinate of ω. Define the Hamming distance by

$$\rho_H(\omega, \omega') := \sum_{k=1}^{m} \mathbf{1}\{\omega_k \neq \omega'_k\}.$$

The vector ω is an m-bit label for the true hypothesis. A decoder must answer m binary questions, and the Hamming distance counts its wrong answers. Assouad's lemma shows that these errors accumulate when the two endpoints of every edge are difficult to distinguish.

Theorem 3 (Assouad's lemma, Theorem 2.12, Tsybakov (2009)) *For each n, let $m = m_n \geq 1$ and let $\{P_{\omega,n} : \omega \in \Omega_m\}$ be probability measures on $(\mathcal{X}^n, \mathcal{A}^{\otimes n})$ associated with the hypotheses. Write $\mathbb{E}_{\omega,n}$ for expectation under $P_{\omega,n}$. Suppose that there exists $\alpha_n \in [0,1)$ such that*

$$\max_{\substack{\omega,\omega'\in\Omega_m \\ \rho_H(\omega,\omega')=1}} V(P_{\omega,n}, P_{\omega',n}) \leq \alpha_n.$$

Then

$$\inf_{\widehat{\omega}_n}\max_{\omega\in\Omega_m} \mathbb{E}_{\omega,n}\left[\rho_H(\widehat{\omega}_n, \omega)\right] \geq \frac{m}{2}(1 - \alpha_n),$$

where the infimum is over all measurable decoders $\widehat{\omega}_n : \mathcal{X}^n \to \Omega_m$.

Proof: Fix a measurable decoder $\widehat{\omega}_n = (\widehat{\omega}_{n,1}, \ldots, \widehat{\omega}_{n,m})$. For a fixed coordinate k, there are only two ways for the decoder to make an error. If $\omega_k = 0$, the error event is $\{\widehat{\omega}_{n,k} = 1\}$, whereas if $\omega_k = 1$, the error event is $\{\widehat{\omega}_{n,k} = 0\}$. We then have

$$\begin{aligned}
\max_{\omega \in \Omega_m} \mathbb{E}_{\omega,n}\left[\rho_H(\widehat{\omega}_n, \omega)\right] &\geq \frac{1}{2^m} \sum_{\omega \in \Omega_m} \mathbb{E}_{\omega,n}\left[\rho_H(\widehat{\omega}_n, \omega)\right] \qquad (14)\\
&= \frac{1}{2^m} \sum_{\omega \in \Omega_m} \mathbb{E}_{\omega,n}\left[\sum_{k=1}^{m} \mathbf{1}\{\widehat{\omega}_{n,k} \neq \omega_k\}\right]\\
&= \frac{1}{2^m} \sum_{k=1}^{m} \sum_{\omega \in \Omega_m} P_{\omega,n}(\widehat{\omega}_{n,k} \neq \omega_k)\\
&= \frac{1}{2^m} \sum_{k=1}^{m} \left\{ \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} P_{\omega,n}(\widehat{\omega}_{n,k} = 1) + \sum_{\substack{\omega' \in \Omega_m \\ \omega'_k = 1}} P_{\omega',n}(\widehat{\omega}_{n,k} = 0) \right\}\\
&= \frac{1}{2^m} \sum_{k=1}^{m} \left\{ \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} P_{\omega,n}(\widehat{\omega}_{n,k} = 1) + \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} P_{\omega^{(k)},n}(\widehat{\omega}_{n,k} = 0) \right\}\\
&= \frac{1}{2^m} \sum_{k=1}^{m} \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} \left\{ P_{\omega,n}(\widehat{\omega}_{n,k} = 1) + P_{\omega^{(k)},n}(\widehat{\omega}_{n,k} = 0) \right\}\\
&= \frac{1}{2^m} \sum_{k=1}^{m} \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} \left\{ 1 + P_{\omega,n}(\widehat{\omega}_{n,k} = 1) - P_{\omega^{(k)},n}(\widehat{\omega}_{n,k} = 1) \right\}\\
&\geq \frac{1}{2^m} \sum_{k=1}^{m} \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} \left\{ 1 - \left| P_{\omega,n}(\widehat{\omega}_{n,k} = 1) - P_{\omega^{(k)},n}(\widehat{\omega}_{n,k} = 1) \right| \right\}\\
&\geq \frac{1}{2^m} \sum_{k=1}^{m} \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} \left\{ 1 - V(P_{\omega,n}, P_{\omega^{(k)},n}) \right\} \qquad (15)\\
&\geq \frac{1}{2^m} \sum_{k=1}^{m} \sum_{\substack{\omega \in \Omega_m \\ \omega_k = 0}} (1 - \alpha_n)\\
&= \frac{m 2^{m-1}}{2^m} (1 - \alpha_n)\\
&= \frac{m}{2}(1 - \alpha_n). \qquad (16)
\end{aligned}$$

Equation (14) bounds the maximum risk by its uniform average over the hypercube. Equation (15) then follows from the definition of total variation applied to the measurable event $\{\widehat{\omega}_{n,k} = 1\}$, and the subsequent inequality uses the assumed bound on every neighboring pair. Finally, there are 2^{m-1} vertices with $\omega_k = 0$ for each of the m coordinates. Hence Equation (16) holds for every measurable decoder, and taking the infimum over $\widehat{\omega}_n$ completes the proof. □

To use Theorem 3 we have to bound our original loss ℓ by the Hamming distance. Let Θ_n be the parameter space. For every $\omega \in \Omega_m$ choose $\theta_{\omega,n} \in \Theta_n$ and set

$$P_{\omega,n} := P^{(n)}_{\theta_{\omega,n}}.$$

Suppose that there exists $c_n > 0$ such that, for every measurable estimator T_n, one can construct a single measurable decoder $\widehat{\omega}_n(T_n) : \mathcal{X}^n \to \Omega_m$, independent of the unknown vertex ω, satisfying

$$\ell\left(\psi_n^{-1} d(T_n, g(\theta_{\omega,n}))\right) \geq c_n \rho_H(\widehat{\omega}_n(T_n), \omega)$$

$P_{\omega,n}$-almost surely for every $\omega \in \Omega_m$. If the edge condition in Theorem 3 holds, then

$$\begin{aligned}\inf_{T_n} \sup_{\theta\in\Theta_n} \mathbb{E}_\theta^{(n)} \left[\ell\left(\psi_n^{-1} d(T_n, g(\theta))\right)\right] &\geq c_n \inf_{T_n} \max_{\omega\in\Omega_m} \mathbb{E}_{\omega,n}\left[\rho_H(\widehat{\omega}_n(T_n), \omega)\right] \\ &\geq c_n \inf_{\widehat{\omega}_n} \max_{\omega\in\Omega_m} \mathbb{E}_{\omega,n}\left[\rho_H(\widehat{\omega}_n, \omega)\right] \\ &\geq c_n \frac{m}{2}(1-\alpha_n).\end{aligned}$$

Example 3 (Gaussian Means Revisited) *Consider again Example 1, where*

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(\theta, I_q), \qquad \theta \in \Theta_n = \mathbb{R}^q,$$

and

$$\psi_n = \sqrt{\frac{q}{n}}, \qquad d(\theta, \theta') = \|\theta - \theta'\|_2, \qquad \ell(u) = u^2.$$

Fix $a > 0$. For every $\omega \in \{0,1\}^q$, define

$$\theta_{\omega,n} := \frac{a}{\sqrt{n}}(2\omega - \mathbf{1}_q), \qquad P_{\omega,n} := P^{(n)}_{\theta_{\omega,n}}.$$

If ω and $\omega^{(k)}$ differ only in coordinate k, then

$$\sqrt{n}\left\|\theta_{\omega,n} - \theta_{\omega^{(k)},n}\right\|_2 = 2a.$$

The Gaussian total-variation calculation in Example 1 therefore gives

$$V(P_{\omega,n}, P_{\omega^{(k)},n}) = 1 - 2\Phi(-a),$$

where Φ is the standard normal distribution function.

For an arbitrary measurable estimator $T_n = (T_{n,1}, \ldots, T_{n,q})$, define

$$\widehat{\omega}_{n,k} := \mathbf{1}\{T_{n,k} \geq 0\}, \qquad k = 1, \ldots, q.$$

Whenever $\widehat{\omega}_{n,k} \neq \omega_k$,

$$|T_{n,k} - \theta_{\omega,n,k}| \geq \frac{a}{\sqrt{n}}.$$

Consequently,

$$\|T_n - \theta_{\omega,n}\|_2^2 \geq \frac{a^2}{n}\rho_H(\widehat{\omega}_n, \omega),$$

and the normalized loss satisfies

$$\ell\left(\psi_n^{-1}\|T_n - \theta_{\omega,n}\|_2\right) = \frac{n}{q}\|T_n - \theta_{\omega,n}\|_2^2 \geq \frac{a^2}{q}\rho_H(\widehat{\omega}_n, \omega).$$

Applying Theorem 3 with $m = q$, $c_n = a^2/q$, and $\alpha_n = 1 - 2\Phi(-a)$ yields

$$\inf_{T_n} \sup_{\theta\in\Theta_n} \mathbb{E}_\theta^{(n)} \left[\ell\left(\psi_n^{-1}\|T_n - \theta\|_2\right)\right] \geq a^2\Phi(-a).$$

Equivalently, on the raw squared-error scale,

$$\inf_{T_n} \sup_{\theta\in\Theta_n} \mathbb{E}_\theta^{(n)} \left[\|T_n - \theta\|_2^2\right] \geq a^2\Phi(-a)\frac{q}{n}.$$

Taking $a = 1$ gives the lower bound $\Phi(-1)q/n$. Together with the upper bound q/n attained by the sample mean, this proves the minimax squared-risk rate q/n. Unlike the Fano argument, the Assouad proof requires neither a Varshamov–Gilbert packing nor an average Kullback–Leibler calculation; it uses the full hypercube and a single neighboring-edge calculation.

The convenience of Assouad's lemma comes with two restrictions. The parameter space must contain a suitable hypercube, and the original loss must dominate ρ_H. This eduction is available only for particular losses and semi-distances. For the indicator loss $\ell(u) = \mathbf{1}\{u \geq A\}$ or the L_∞-distance, coordinate errors do not accumulate; the natural comparison is therefore with normalized Hamming loss and loses the factor m. Consequently, the standard Assouad reduction does not generally yield a useful dimension-dependent lower bound for these criteria (Tsybakov, 2009, p. 120).

3.2.3 The Method of Two Fuzzy Hypotheses

Le Cam's two-point method restricts attention to two fixed parameters. The method of two fuzzy hypotheses instead places a probability measure on each hypothesis and compares the resulting mixture distributions. This additional flexibility is useful when averaging produces distributions that are difficult to distinguish even though the individual distributions in their supports are not close.

In this subsection, suppose that $\mathcal{G} = \mathbb{R}$ and $d(u, v) = |u - v|$. For each n, let $\mu_{0,n}$ and $\mu_{1,n}$ be probability measures on Θ, and define the induced mixture distributions

$$\overline{P}_{i,n}(B) := \int_{\Theta} P_{\theta}^{(n)}(B)\, \mu_{i,n}(d\theta), \qquad B \in \mathcal{A}^{\otimes n}, \quad i \in \{0, 1\}.$$

Theorem 4 (Two fuzzy hypotheses, Theorem 2.15, Tsybakov (2009)) *Suppose that there exist $c_n \in \mathbb{R}$, $A_n > 0$ with $\ell(A_n) > 0$, and $\beta_{0,n}, \beta_{1,n} \in [0, 1)$ such that*

1. $\mu_{0,n}\big(\{\theta \in \Theta : g(\theta) \le c_n\}\big) \ge 1 - \beta_{0,n}$,
2. $\mu_{1,n}\big(\{\theta \in \Theta : g(\theta) \ge c_n + 2A_n\psi_n\}\big) \ge 1 - \beta_{1,n}$.

Then

$$\inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}_{\theta}^{(n)} \left[\ell\left(\psi_n^{-1} |T_n - g(\theta)|\right) \right] \ge \frac{\ell(A_n)}{2} \left[1 - V\big(\overline{P}_{0,n}, \overline{P}_{1,n}\big) - \beta_{0,n} - \beta_{1,n} \right].$$

Theorem 4 balances separation of the estimand under the two priors against closeness of the induced mixture distributions. The bound is informative when

$$V\big(\overline{P}_{0,n}, \overline{P}_{1,n}\big) + \beta_{0,n} + \beta_{1,n} < 1.$$

If both priors are Dirac measures and $\beta_{0,n} = \beta_{1,n} = 0$, the result reduces to the two-point bound in Theorem 1.

The preceding result requires the estimand to lie on opposite sides of a fixed threshold with high probability under the two priors. The constrained-risk inequality of Cai and Low (2011) replaces this condition by separation of the prior means and explicitly accounts for variation of the estimand under the first prior.

For $i \in \{0, 1\}$, define

$$m_{i,n} := \int_{\Theta} g(\theta)\, \mu_{i,n}(d\theta), \qquad v_{i,n}^2 := \int_{\Theta} \big(g(\theta) - m_{i,n}\big)^2\, \mu_{i,n}(d\theta).$$

Suppose that $\overline{P}_{1,n} \ll \overline{P}_{0,n}$ and define

$$I_n := \left[\int_{\mathcal{X}^n} \left(\frac{d\overline{P}_{1,n}}{d\overline{P}_{0,n}} - 1 \right)^2 d\overline{P}_{0,n} \right]^{1/2} = \sqrt{\chi^2\big(\overline{P}_{1,n}, \overline{P}_{0,n}\big)}.$$

Theorem 5 (Composite constrained-risk inequality, Corollary 1, Cai and Low (2011)) *Suppose that $I_n < \infty$ and*

$$|m_{1,n} - m_{0,n}| > v_{0,n} I_n.$$

Then

$$\inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}_{\theta}^{(n)} \left[\big(T_n - g(\theta)\big)^2 \right] \ge \frac{\big(|m_{1,n} - m_{0,n}| - v_{0,n} I_n\big)^2}{(I_n + 2)^2}.$$

Unlike Theorem 4, Theorem 5 allows the prior supports to overlap and does not require uniform or high-probability separation of $g(\theta)$. The effective separation is instead $|m_{1,n} - m_{0,n}| - v_{0,n} I_n$. In particular, if

$$|m_{1,n} - m_{0,n}| - v_{0,n} I_n \gtrsim \psi_n \qquad \text{and} \qquad I_n \lesssim 1,$$

then the theorem yields a minimax squared-risk lower bound of order ψ_n^2.

A Distances Between Probability Measures

Let P and Q be two probability measures on $\mathcal{X}^{(n)}$. Let ν be a measure dominating both P and Q, and write

$$p := \frac{dP}{d\nu}, \qquad q := \frac{dQ}{d\nu}.$$

Such a measure always exists, for example, one may take $\nu = P + Q$. The quantities below do not depend on the choice of ν.

A.1 Total Variation Distance

The *total variation distance* between P and Q is

$$V(P, Q) := \sup_{B \in \mathcal{A}^{(n)}} |P(B) - Q(B)|.$$

Properties of the total variation distance.

1. It is a distance and satisfies $0 \le V(P, Q) \le 1$.
2. It can be calculated directly from the densities using

$$V(P, Q) = \frac{1}{2} \int_{\mathcal{X}^{(n)}} |p - q| \, d\nu = 1 - \int_{\mathcal{X}^{(n)}} \min\{p, q\} \, d\nu.$$

A.2 Hellinger Distance

The *Hellinger distance* between P and Q is

$$H(P, Q) := \left(\int_{\mathcal{X}^{(n)}} (\sqrt{p} - \sqrt{q})^2 \, d\nu \right)^{1/2}.$$

Properties of the Hellinger distance.

1. It is a distance and satisfies $0 \le H^2(P, Q) \le 2$.
2. $H^2(P, Q) = 2 \left(1 - \int_{\mathcal{X}^{(n)}} \sqrt{pq} \, d\nu\right)$.
3. If $P = \bigotimes_{i=1}^n P_i$ and $Q = \bigotimes_{i=1}^n Q_i$ are product measures, then

$$H^2(P, Q) = 2 \left[1 - \prod_{i=1}^n \left(1 - \frac{H^2(P_i, Q_i)}{2} \right) \right],$$

   which reduces the calculation to the component distributions.

A.3 Kullback–Leibler Divergence

The *Kullback–Leibler divergence* of P from Q is

$$K(P, Q) := \begin{cases} \displaystyle\int_{\mathcal{X}^{(n)}} \log\left(\frac{dP}{dQ}\right) dP, & P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Properties of the Kullback–Leibler divergence.

1. If $P \ll Q$, it can be calculated as

$$K(P, Q) = \int_{\{pq > 0\}} p \log\left(\frac{p}{q}\right) d\nu.$$

2. It is nonnegative and equals zero if and only if $P = Q$, but it is not symmetric and hence is not a distance.

3. For product measures,
$$K\left(\bigotimes_{i=1}^{n} P_i, \bigotimes_{i=1}^{n} Q_i\right) = \sum_{i=1}^{n} K(P_i, Q_i).$$
In particular, in the i.i.d. case,
$$K(P_1^{\otimes n}, Q_1^{\otimes n}) = nK(P_1, Q_1).$$

A.4 Chi-Squared Divergence

The χ^2-*divergence* of P from Q is
$$\chi^2(P, Q) := \begin{cases} \displaystyle\int_{\mathcal{X}^{(n)}} \left(\frac{dP}{dQ} - 1\right)^2 dQ, & P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Properties of the χ^2-divergence.

1. When $P \ll Q$, it admits the convenient expression
$$\chi^2(P, Q) = \int_{\{pq>0\}} \frac{p^2}{q}\, d\nu - 1.$$

2. It is nonnegative and equals zero if and only if $P = Q$, but it is not symmetric and hence is not a distance.

3. For product measures,
$$1 + \chi^2\left(\bigotimes_{i=1}^{n} P_i, \bigotimes_{i=1}^{n} Q_i\right) = \prod_{i=1}^{n} \left(1 + \chi^2(P_i, Q_i)\right).$$
In particular, in the i.i.d. case,
$$\chi^2(P_1^{\otimes n}, Q_1^{\otimes n}) = \left(1 + \chi^2(P_1, Q_1)\right)^n - 1.$$

B Information-Theoretic Quantities

We briefly introduce the information-theoretic quantities used throughout this section. All logarithms are natural, and we adopt the convention $0 \log 0 = 0$. Let U take values in a finite set $\mathcal{U}$, and let V be an arbitrary random variable.

B.1 Entropy

The *entropy* of U is
$$\mathsf{H}(U) := -\sum_{u \in \mathcal{U}} \mathbb{P}(U = u) \log \mathbb{P}(U = u).$$
It measures the uncertainty about U before any additional information is observed. Its main properties are

1. $0 \le \mathsf{H}(U) \le \log |\mathcal{U}|$;

2. $\mathsf{H}(U) = 0$ if U is constant;

3. $\mathsf{H}(U) = \log |\mathcal{U}|$ if U is uniformly distributed.

B.2 Conditional Entropy

The *conditional entropy* of U given Z is

$$\mathsf{H}(U \mid Z) := -\mathbb{E}\left[\sum_{u \in \mathcal{U}} \mathbb{P}(U = u \mid Z) \log \mathbb{P}(U = u \mid Z)\right].$$

It measures the uncertainty about U that remains after observing Z. The main properties used below are

1. Conditioning reduces entropy: $\mathsf{H}(U \mid Z) \leq \mathsf{H}(U)$;
2. Chain Rule: $\mathsf{H}(U, W \mid Z) = \mathsf{H}(W \mid Z) + \mathsf{H}(U \mid W, Z)$.

B.3 Mutual Information

The *mutual information* between U and Z is

$$\mathsf{I}(U; Z) := \mathsf{H}(U) - \mathsf{H}(U \mid Z).$$

It measures the reduction in uncertainty about U obtained by observing Z. Equivalently,

$$\mathsf{I}(U; Z) = K(P_{U,Z}, P_U \otimes P_Z) = \sum_{u \in \mathcal{U}} \mathbb{P}(U = u) K(P_{Z|U=u}, P_Z).$$

Consequently,

1. $\mathsf{I}(U; Z) \geq 0$;
2. $\mathsf{I}(U; Z) = 0$ if and only if U and Z are independent;
3. $\mathsf{I}(U; Z) \leq \mathsf{H}(U)$.

References

T. Tony Cai and Mark G. Low. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2):1012–1041, 2011. doi: 10.1214/10-AOS849.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 1 edition, 2009. ISBN 978-0-387-79052-7. doi: 10.1007/b13794.