7

Slow Rates of Convergence

In this chapter we consider the general pattern recognition problem: Given the observation X and the training data $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ of independent identically distributed random variable pairs, we estimate the label Y by the decision

$$g_n(X) = g_n(X, D_n).$$

The error probability is

$$L_n = \mathbf{P}\{Y \neq g_n(X) | D_n\}.$$

Obviously, the average error probability $\mathbf{E}L_n = \mathbf{P}\{Y \neq g_n(X)\}$ is completely determined by the distribution of the pair (X, Y), and the classifier g_n. We have seen in Chapter 6 that there exist classification rules such as the cubic histogram rule with properly chosen cube sizes such that $\lim_{n\to\infty} \mathbf{E}L_n = L^*$ for all possible distributions. The next question is whether there are classification rules with $\mathbf{E}L_n$ tending to the Bayes risk at a specified rate for all distributions. Disappointingly, such rules do not exist.

7.1 Finite Training Sequence

The first negative result shows that for any classification rule and for any *fixed* n, there exists a distribution such that the difference between the error probability of the rule and L^* is larger than $1/4$. To explain this, note that for fixed n, we can find a sufficiently complex distribution for which the sample size n is hopelessly small.

Theorem 7.1. (DEVROYE (1982B)). *Let $\epsilon > 0$ be an arbitrarily small number. For any integer n and classification rule g_n, there exists a distribution of (X, Y) with Bayes risk $L^* = 0$ such that*

$$\mathbf{E}L_n \geq 1/2 - \epsilon.$$

PROOF. First we construct a family of distributions of (X, Y). Then we show that the error probability of any classifier is large for at least one member of the family. For every member of the family, X is uniformly distributed on the set $\{1, 2, \ldots, K\}$ of positive integers

$$p_i = \mathbf{P}\{X = i\} = \begin{cases} 1/K & \text{if } i \in \{1, \ldots, K\} \\ 0 & \text{otherwise,} \end{cases}$$

where K is a large integer specified later. Now, the family of distributions of (X, Y) is parameterized by a number $b \in [0, 1)$, that is, every b determines a distribution as follows. Let $b \in [0, 1)$ have binary expansion $b = 0.b_0b_1b_2\ldots$, and define $Y = b_X$. As the label Y is a function of X, there exists a perfect decision, and thus $L^* = 0$. We show that for any decision rule g_n there is a b such that if $Y = b_X$, then g_n has very poor performance. Denote the average error probability corresponding to the distribution determined by b, by $R_n(b) = \mathbf{E}L_n$.

The proof of the existence of a bad distribution is based on the so-called probabilistic method. Here the key trick is the randomization of b. Define a random variable B which is uniformly distributed in $[0, 1)$ and independent of X and $X_1, \ldots, X_n$. Then we may compute the expected value of the random variable $R_n(B)$. Since for any decision rule g_n,

$$\sup_{b \in [0,1)} R_n(b) \geq \mathbf{E}\{R_n(B)\},$$

a lower bound for $\mathbf{E}\{R_n(B)\}$ proves the existence of a $b \in [0, 1)$ whose corresponding error probability exceeds the lower bound.

Since B is uniformly distributed in $[0, 1)$, its binary extension $B = 0.B_1B_2\ldots$ is a sequence of independent binary random variables with $\mathbf{P}\{B_i = 0\} = \mathbf{P}\{B_i = 1\} = 1/2$. But

$$\begin{aligned}
&\mathbf{E}\{R_n(B)\} \\
&\quad = \mathbf{P}\{g_n(X, D_n) \neq Y\} \\
&\quad = \mathbf{P}\{g_n(X, D_n) \neq B_X\} \\
&\quad = \mathbf{P}\left\{g_n(X, X_1, B_{X_1}, \ldots, X_n, B_{X_n}) \neq B_X\right\} \\
&\quad = \mathbf{E}\left\{\mathbf{P}\left\{g_n(X, X_1, B_{X_1}, \ldots, X_n, B_{X_n}) \neq B_X \,\middle|\, X, X_1, \ldots, X_n\right\}\right\} \\
&\quad \geq \frac{1}{2}\mathbf{P}\{X \neq X_1, X \neq X_2, \ldots, X \neq X_n\},
\end{aligned}$$

since if $X \neq X_i$ for all $i = 1, 2, \ldots, n$, then given $X, X_1, \ldots, X_n$, $Y = B_X$ is conditionally independent of $g_n(X, D_n)$ and Y takes values 0 and 1 with probability

1/2. But clearly,

$$\mathbf{P}\{X \neq X_1, X \neq X_2, \ldots, X \neq X_n | X\} = \mathbf{P}\{X \neq X_1 | X\}^n = (1 - 1/K)^n.$$

In summary,

$$\sup_{b \in [0,1)} R_n(b) \geq \frac{1}{2}(1 - 1/K)^n.$$

The lower bound tends to 1/2 as $K \to \infty$. □

Theorem 7.1 states that even though we have rules that are universally consistent, that is, they *asymptotically* provide the optimal performance for any distribution, their *finite sample* performance is *always* extremely bad for some distributions. This means that no classifier guarantees that with a sample size of (say) $n = 10^8$ we get within 1/4 of the Bayes error probability for all distributions. However, as the bad distribution depends upon n, Theorem 7.1 does not allow us to conclude that there is one distribution for which the error probability is more than $L^* + 1/4$ for all n. Indeed, that would contradict the very existence of universally consistent rules.

7.2 Slow Rates

The next question is whether a certain universal rate of convergence to L^* is achievable for some classifier. For example, Theorem 7.1 does not exclude the existence of a classifier such that for every n, $\mathbf{E}L_n - L^* \leq c/n$ for all distributions, for some constant c depending upon the actual distribution. The next negative result is that this cannot be the case. Theorem 7.2 below states that the error probability $\mathbf{E}L_n$ of any classifier is larger than (say) $L^* + c/(\log\log\log n)$ *for every* n for some distribution, even if c depends on the distribution. (This can be seen by considering that by Theorem 7.2, there exists a distribution of (X, Y) such that $\mathbf{E}L_n \geq L^* + 1/\sqrt{\log\log\log n}$ for every n.) Moreover, there is no sequence of numbers a_n converging to zero such that there is a classification rule with error probability below L^* plus a_n for all distributions.

Thus, in practice, no classifier assures us that its error probability is close to L^*, unless the actual distribution is known to be a member of a restricted class of distributions. Now, it is easily seen that in the proof of both theorems we could take X to have uniform distribution on [0, 1], or any other density (see Problem 7.2). Therefore, putting restrictions on the distribution of X alone does not suffice to obtain rate-of-convergence results. For such results, one needs conditions on the a posteriori probability $\eta(x)$ as well. However, if only training data give information about the joint distribution, then theorems with extra conditions on the distribution have little practical value, as it is impossible to detect whether, for example, the a posteriori probability $\eta(x)$ is twice differentiable or not.

Now, the situation may look hopeless, but this is not so. Simply put, the Bayes error is too difficult a target to shoot at.

Weaker versions of Theorem 7.2 appeared earlier in the literature. First Cover (1968b) showed that for any sequence of classification rules, for sequences $\{a_n\}$ converging to zero at arbitrarily slow algebraic rates (i.e., as $1/n^\delta$ for arbitrarily small $\delta > 0$), there exists a distribution such that $\mathbf{E}L_n \geq L^* + a_n$ infinitely often. Devroye (1982b) strengthened Cover's result allowing sequences tending to zero arbitrarily slowly. The next result asserts that $\mathbf{E}L_n > L^* + a_n$ for *every* n.

Theorem 7.2. *Let $\{a_n\}$ be a sequence of positive numbers converging to zero with $1/16 \geq a_1 \geq a_2 \geq \ldots$. For every sequence of classification rules, there exists a distribution of (X, Y) with $L^* = 0$, such that*

$$\mathbf{E}L_n \geq a_n$$

for all n.

This result shows that universally good classification rules do not exist. Rate of convergence studies for particular rules must necessarily be accompanied by conditions on (X, Y). That these conditions too are necessarily restrictive follows from examples suggested in Problem 7.2. Under certain regularity conditions it is possible to obtain upper bounds for the rates of convergence for the probability of error of certain rules to L^*. Then it is natural to ask what the fastest achievable rate is for the given class of distributions. A theory for regression function estimation was worked out by Stone (1982). Related results for classification were obtained by Marron (1983). In the proof of Theorem 7.2 we will need the following simple lemma:

Lemma 7.1. *For any monotone decreasing sequence $\{a_n\}$ of positive numbers converging to zero with $a_1 \leq 1/16$, a probability distribution $(p_1, p_2, \ldots)$ may be found such that $p_1 \geq p_2 \geq \ldots$, and for all n*

$$\sum_{i=n+1}^{\infty} p_i \geq \max\left(8a_n, 32np_{n+1}\right).$$

PROOF. It suffices to look for p_i's such that

$$\sum_{i=n+1}^{\infty} p_i \geq \max\left(8a_n, 32np_n\right).$$

These conditions are easily satisfied. For positive integers $u < v$, define the function $H(v, u) = \sum_{i=u}^{v-1} 1/i$. First we find a sequence $1 = n_1 < n_2 < \ldots$ of integers with the following properties:

(a) $H(n_{k+1}, n_k)$ is monotonically increasing,

(b) $H(n_2, n_1) \geq 32$,

(c) $8a_{n_k} \leq 1/2^k$ for all $k \geq 1$.

Note that (c) may only be satisfied if $a_{n_1} = a_1 \leq 1/16$. To this end, define constants $c_1, c_2, \ldots$ by

$$c_k = \frac{32}{2^k H(n_{k+1}, n_k)}, \quad k \geq 1,$$

so that the c_k's are decreasing in k, and

$$\frac{1}{32} \sum_{k=1}^{\infty} c_k H(n_{k+1}, n_k) = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1.$$

For $n \in [n_k, n_{k+1})$, we define $p_n = c_k/(32n)$. We claim that these numbers have the required properties. Indeed, $\{p_n\}$ is decreasing, and

$$\sum_{n=1}^{\infty} p_n = \sum_{k=1}^{\infty} \frac{c_k}{32} H(n_{k+1}, n_k) = 1.$$

Finally, if $n \in [n_k, n_{k+1})$, then

$$\sum_{i=n+1}^{\infty} p_i \geq \sum_{j=k+1}^{\infty} \frac{c_j}{32} H(n_{j+1}, n_j) = \sum_{j=k+1}^{\infty} \frac{1}{2^j} = \frac{1}{2^k}.$$

Clearly, on the one hand, by the monotonicity of $H(n_{k+1}, n_k)$, $1/2^k \geq c_k = 32np_n$. On the other hand, $1/2^k \geq 8a_{n_k} \geq 8a_n$. This concludes the proof. □

PROOF OF THEOREM 7.2. We introduce some notation. Let $b = 0.b_1b_2b_3\ldots$ be a real number on [0, 1] with the shown binary expansion, and let B be a random variable uniformly distributed on [0, 1] with expansion $B = 0.B_1B_2B_3\ldots$. Let us restrict ourselves to a random variable X with

$$\mathbf{P}\{X = i\} = p_i, \quad i \geq 1,$$

where $p_1 \geq p_2 \geq \ldots > 0$, and $\sum_{i=n+1}^{\infty} p_i \geq \max(8a_n, 32np_{n+1})$ for every n. That such p_i's exist follows from Lemma 7.1. Set $Y = b_X$. As Y is a function of X, we see that $L^* = 0$. Each $b \in [0, 1)$ however describes a different distribution. With b replaced by B we have a random distribution. Introduce the short notation $\Delta_n = ((X_1, B_{X_1}), \ldots, (X_n, B_{X_n}))$, and define $G_{ni} = g_n(i, \Delta_n)$. If $L_n(B)$ denotes the probability of error $\mathbf{P}\{g_n(X, \Delta_n) \neq Y | B, X_1, \ldots, X_n\}$ for the random distribution, then we note that we may write

$$L_n(B) = \sum_{i=1}^{\infty} p_i I_{\{G_{ni} \neq B_i\}}.$$

If $L_n(b)$ is the probability of error for a distribution parametrized by b, then

$$\begin{aligned}\sup_b \inf_n \mathbf{E}\frac{L_n(b)}{2a_n} &\geq \sup_b \mathbf{E}\left\{\inf_n \frac{L_n(b)}{2a_n}\right\} \\ &\geq \mathbf{E}\left\{\inf_n \frac{L_n(B)}{2a_n}\right\} \\ &= \mathbf{E}\left\{\mathbf{E}\left\{\inf_n \frac{L_n(B)}{2a_n}\middle| X_1, X_2, \ldots\right\}\right\}.\end{aligned}$$

We consider only the conditional expectation for now. We have

$$\begin{aligned}&\mathbf{E}\left\{\inf_n \frac{L_n(B)}{2a_n}\middle| X_1, X_2, \ldots\right\} \\ &\geq \mathbf{P}\left\{\bigcap_{n=1}^{\infty}\{L_n(B) \geq 2a_n\}\middle| X_1, X_2, \ldots\right\} \\ &\geq 1 - \sum_{n=1}^{\infty}\mathbf{P}\{L_n(B) < 2a_n | X_1, X_2, \ldots\} \\ &= 1 - \sum_{n=1}^{\infty}\mathbf{P}\{L_n(B) < 2a_n | X_1, X_2, \ldots, X_n\} \\ &= 1 - \sum_{n=1}^{\infty}\mathbf{E}\{\mathbf{P}\{L_n(B) < 2a_n | \Delta_n\} | X_1, X_2, \ldots, X_n\}.\end{aligned}$$

We bound the conditional probabilities inside the sum:

$$\begin{aligned}&\mathbf{P}\{L_n(B) < 2a_n | \Delta_n\} \\ &\leq \mathbf{P}\left\{\sum_{i \notin \{X_1,\ldots,X_n\}} p_i I_{\{G_{ni} \neq B_i\}} < 2a_n \middle| \Delta_n\right\} \\ &\quad \text{(and, noting that } G_{ni}, X_1, \ldots, X_n \\ &\quad \text{are all functions of } \Delta_n, \text{ we have:)} \\ &= \mathbf{P}\left\{\sum_{i \notin \{X_1,\ldots,X_n\}} p_i I_{\{B_i=1\}} < 2a_n \middle| \Delta_n\right\} \\ &\leq \mathbf{P}\left\{\sum_{i=n+1}^{\infty} p_i I_{\{B_i=1\}} < 2a_n\right\} \\ &\quad \text{(since the } p_i\text{'s are decreasing, by stochastic dominance)} \\ &= \mathbf{P}\left\{\sum_{i=n+1}^{\infty} p_i B_i < 2a_n\right\}.\end{aligned}$$

Now everything boils down to bounding these probabilities from above. We proceed by *Chernoff's bounding technique*. The idea is the following: For any random variable X, and $s > 0$, by Markov's inequality,

$$\mathbf{P}\{X \geq \epsilon\} = \mathbf{P}\{e^{sX} \geq e^{s\epsilon}\} \leq \frac{\mathbf{E}\left\{e^{sX}\right\}}{e^{s\epsilon}}.$$

By cleverly choosing s one can often obtain very sharp bounds. For more discussion and examples of Chernoff's method, refer to Chapter 8. In our case,

$$\begin{aligned}
&\mathbf{P}\left\{\sum_{i=n+1}^{\infty} p_i B_i < 2a_n\right\} \\
&\leq \mathbf{E}\left\{e^{2sa_n - s\sum_{i=n+1}^{\infty} p_i B_i}\right\} \\
&= e^{2sa_n} \prod_{i=n+1}^{\infty} \left(\frac{1}{2} + \frac{1}{2}e^{-sp_i}\right) \\
&\leq e^{2sa_n} \prod_{i=n+1}^{\infty} \frac{1}{2}\left(2 - sp_i + \frac{s^2 p_i^2}{2}\right) \\
&\quad \text{(since } e^{-x} \leq 1 - x + x^2/2 \text{ for } x \geq 0\text{)} \\
&\leq \exp\left(2sa_n + \sum_{i=n+1}^{\infty}\left(-\frac{sp_i}{2} + \frac{s^2 p_i^2}{4}\right)\right) \\
&\quad \text{(since } 1 - x \leq e^{-x}\text{)} \\
&\leq \exp\left(2sa_n - \frac{s\Sigma}{2} + \frac{s^2 p_{n+1}\Sigma}{4}\right) \\
&\quad \text{(where } \Sigma = \textstyle\sum_{i=n+1}^{\infty} p_i\text{)} \\
&= \exp\left(-\frac{1}{4}\frac{(4a_n - \Sigma)^2}{\Sigma p_{n+1}}\right) \\
&\quad \text{(by taking } s = \tfrac{\Sigma - 4a_n}{p_{n+1}\Sigma}\text{, and the fact that } \Sigma > 4a_n\text{)} \\
&\leq \exp\left(-\frac{1}{16}\frac{\Sigma}{p_{n+1}}\right) \quad \text{(since } \Sigma \geq 8a_n\text{)} \\
&\leq e^{-2n} \quad \text{(since } \Sigma \geq 32 p_{n+1} n\text{)}.
\end{aligned}$$

Thus, we conclude that

$$\sup_b \inf_n \mathbf{E}\frac{L_n(b)}{2a_n} \geq 1 - \sum_{n=1}^{\infty} e^{-2n} = \frac{e^2 - 2}{e^2 - 1} > \frac{1}{2},$$

so that there exists a b for which $\mathbf{E}L_n(b) \geq a_n$ for all n. □

Problems and Exercises

PROBLEM 7.1. Extend Theorem 7.2 for distributions with $0 < L^* < 1/2$: show that if a_n is a sequence of positive numbers as in Theorem 7.2, then for any classification rule there is a distribution such that $\mathbf{E}L_n - L^* \geq a_n$ for every n for which $L^* + a_n < 1/2$.

PROBLEM 7.2. Prove Theorems 7.1 and 7.2, under one of the following additional assumptions, which make the case that one will need very restrictive conditions indeed to study rates of convergence.

(1) X has a uniform density on $[0, 1)$.
(2) X has a uniform density on $[0, 1)$ and η is infinitely many times continuously differentiable on $[0, 1)$.
(3) η is unimodal in $x \in \mathcal{R}^2$, that is, $\eta(\lambda x)$ decreases as $\lambda > 0$ increases for any $x \in \mathcal{R}^2$.
(4) η is $\{0, 1\}$-valued, X is $\mathcal{R}^2$-valued, and the set $\{x : \eta(x) = 1\}$ is a compact convex set containing the origin.

PROBLEM 7.3. THERE IS NO SUPER-CLASSIFIER. Show that for every sequence of classification rules $\{g_n\}$ there is a universally consistent sequence of rules $\{g'_n\}$, such that for some distribution of (X, Y),

$$\mathbf{P}\{g_n(X) \neq Y\} > \mathbf{P}\{g'_n(X) \neq Y\}$$

for all n.

PROBLEM 7.4. The next two exercises are intended to demonstrate that the weaponry of pattern recognition can often be successfully used for attacking other statistical problems. For example, a consequence of Theorem 7.2 is that estimating infinite discrete distributions is hard. Consider the problem of estimating a distribution $(p_1, p_2, \ldots)$ on the positive integers $\{1, 2, 3, \ldots\}$ from a sample $X_1, \ldots, X_n$ of i.i.d. random variables with $\mathbf{P}\{X_1 = i\} = p_i$, $i \geq 1$. Show that for any decreasing sequence $\{a_n\}$ of positive numbers converging to zero with $a_1 \leq 1/16$, and any estimate $\{p_{i,n}\}$, there exists a distribution such that

$$\mathbf{E}\left\{\sum_{i=1}^{\infty} |p_i - p_{i,n}|\right\} \geq a_n.$$

HINT: Consider a classification problem with $L^* = 0$, $\mathbf{P}\{Y = 0\} = 1/2$, and X concentrated on $\{1, 2, \ldots\}$. Assume that the class-conditional probabilities $p_i^{(0)} = \mathbf{P}\{X = i | Y = 0\}$ and $p_i^{(1)} = \mathbf{P}\{X = i | Y = 1\}$ are estimated from two i.i.d. samples $X_1^{(0)}, \ldots, X_n^{(0)}$ and $X_1^{(1)}, \ldots, X_n^{(1)}$, distributed according to $\{p_i^{(0)}\}$ and $\{p_i^{(1)}\}$, respectively. Use Theorem 2.3 to show that for the classification rule obtained from these estimates in a natural way,

$$\mathbf{E}L_n \leq \frac{1}{2}\mathbf{E}\left\{\sum_{i=1}^{\infty} |p_i^{(0)} - p_{i,n}^{(0)}| + \sum_{i=1}^{\infty} |p_i^{(1)} - p_{i,n}^{(1)}|\right\},$$

therefore the lower bound of Theorem 7.2 can be applied.

PROBLEM 7.5. A similar slow-rate result appears in density estimation. Consider the problem of estimating a density f on $\mathcal{R}$, from an i.i.d. sample $X_1, \ldots, X_n$ having density f. Show that for any decreasing sequence $\{a_n\}$ of positive numbers converging to zero with $a_1 \leq 1/16$, and any density estimate f_n, there exists a distribution such that

$$\mathbf{E}\left\{\int |f(x) - f_n(x)|dx\right\} \geq a_n.$$

This result was proved by Birgé (1986) using a different—and in our view much more complicated—argument. HINT: Put $p_i = \int_i^{i+1} f(x)dx$ and $p_{i,n} = \int_i^{i+1} f_n(x)dx$ and apply Problem 7.4.